General Aviation Pilot Situation Assessment and Decision-Making During Flights in Deteriorating Visibility Conditions

Ulf Ahlstrom, AvMet Applications, Inc.
Nicole Racine, DSoft Technology, Engineering, and Analysis, Inc.
Eamon Caddigan, DSoft Technology, Engineering, and Analysis, Inc.
Kenneth Schulz, TASC: An Engility Company
Kevin Hallman, TASC: An Engility Company

August 2019

Technical Report

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof. The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report. This document does not constitute Federal Aviation Administration (FAA) certification policy. Consult your local FAA aircraft certification office as to its use.

This report is available at the FAA William J. Hughes Technical Center's full-text Technical Reports Web site: http://actlibrary.tc.faa.gov in Adobe® Acrobat® portable document format (PDF).

1. Report No. DOT/FAA/TC-19/32	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle General Aviation Pilot Situation Assessment and Decision-Making During Flights in Deteriorating Visibility Conditions		5. Report Date August 2019	
		6. Performing Organization Code ANG-E25	
7. Author(s) Ulf Ahlstrom, FAA Human Factors Branch Nicole Racine, DSoft Technology, Engineering, and Analysis, Inc. Eamon Caddigan, DSoft Technology, Engineering, and Analysis, Inc. Ken Schulz, TASC, an Engility Company Kevin Hallman, TASC, an Engility Company		8. Performing Organization Report No. DOT/FAA/TC-19/32	
9. Performing Organization Name and Address Federal Aviation Administration Human Factors Branch William J. Hughes Technical Center Atlantic City International Airport, NJ 08405		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No.	
12. Sponsoring Agency Name and Address Federal Aviation Administration 800 Independence Avenue, S.W. Washington, DC 20591		13. Type of Report and Period Covered Technical Report	
		14. Sponsoring Agency Code ANG-C61	
15. Supplementary Notes			

16. Abstract

One of the most dangerous things a General Aviation (GA) pilot who is flying Visual Flight Rules (VFR) flights can do is fly into Instrument Meteorological Conditions (IMC) (bad weather areas where they can no longer fly solely by means of visual references), yet it happens too often, despite access to weather information. This paper describes two human-in-the-loop cockpit simulations to assess whether enhancing the salience of the weather symbols or using a decision support tool (referred to as active reminders) impact pilot decision-making and behavior during flights in deteriorating weather conditions. Ninety-three private GA pilots participated in Simulation 1 and 95 participated in Simulation 2. Pilots were between the ages of 18-83. In Simulation 1, participants were randomly allocated to one of two enhanced weather presentations. In Simulation 1, our results show no difference between pilot use of frequency-tuned salience analysis and enhanced METAR symbols (i.e., triangles and circles) with detection rates of 52% and 62%. The detection rate for METAR triangles in the study is higher than previous research; however, the effect of increasing the METAR symbol salience on change-detection performance was only about 12%, which is not enough to guarantee an unfailing detection by pilots. In Simulation 2, participants were randomly allocated to one of two groups, one of which provided a blue line as an active reminder of forecasted visibility conditions and the distance from the aircraft to hazardous precipitation. In Simulation 2, the results shows a clear benefit of the active reminder display with a credibly higher weather situation awareness for the experimental group compared to the control group (no active reminder/blue line). We also found that, on average, pilots underestimate the "out-the-window" visibility for a range of simulated visibilities. This reveals a gap in pilot situational assessments and undermines a high weather situational awareness during flights in deteriorating visibility. The failure to assess current visibility conditions increases the odds of VFR-into-IMC flights.

17. Key Words change blindness, change detection, decision making, enhanced symbols, General Aviation, METAR symbols, pilot, visibility, weather		18. Distribution Statement This document is available to the public through the National Technical Information Service, Alexandria, Virginia, 22312. A copy is retained for reference at the William J. Hughes Technical Center Library.	
19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. of Pages 102	22. Price
Form DOT F 1700.7 (8-72)	Reproduction of completed page authorized		

Table of Contents

List of Illustrations

Figures   Page

Tables   Page

Acknowledgments

Executive Summary

The goal of the study, conducted under the Weather Technology in the Cockpit (WITC) program, was to develop Minimum Weather Service rendering recommendations to resolve/reduce the previously identified gap of change blindness and reduce/resolve associated safety risks. Researchers used the Federal Aviation Administration (FAA) William J. Hughes Technical Center Cockpit Simulator Facility to perform two cockpit simulations. The cockpit simulations assessed pilot decision-making and behavior during flights in deteriorating weather conditions.

The study consisted of two between-subjects flight simulations. In Simulation1, participants were randomly allocated to one of two enhanced weather presentations that assessed pilots' sensitivity to changes in METAR symbols during flight. In Simulation 2, participants were randomly allocated to one of two groups, only one of which used a portable weather application that provided an Active Reminder (AR) of forecasted visibility conditions and the distance from the aircraft to hazardous precipitation. Simulation 2 assessed the effect on pilot flying behavior and decision-making from manipulations of visibility along the route of flight.

In Simulation 1, we subjected our METAR symbols to a frequency-tuned salience analysis and enhanced our METAR symbols by increasing the symbol salience. Our results show no difference in pilot use of METAR symbols (i.e., triangles and circles) with visual flight rules (VFR) to instrument flight rules (IFR) detection rates of 52% and 62%. The detection rate for METAR triangles in the study is higher than what was found by Ahlstrom and Suss (2014); however, the effect of increasing the METAR symbol salience on change-detection performance was only approximately 12%. Therefore, considering the small performance gains, we believe that METAR color changes (i.e., salience differences) are not enough to guarantee an unfailing detection by pilots.

In Simulation 2, we used a blue line on the display to notify pilots they were 20 nmi away from 30 dBZ precipitation cells or 20 nmi away from forecasted areas of 3 nmi or lower visibility. The result shows clear benefits of the AR display with a credibly higher Weather Situation Awareness (WSA) for the experimental group compared to the control group with no AR. This was expressed as a clearer pilot understanding of the closest point of approach to hazardous precipitation cells and credibly fewer entries into one nmi visibility zones (i.e., Instrument Meteorological Conditions [IMC]). When asked about the safety distance, 91% of all study pilots agreed with the current recommendation to avoid hazardous storm cells by at least 20 statute miles. However, when asked if they flew closer than 20 miles during the simulation, only between 33% (Control) and 63% (Experimental) of pilots provided a correct answer. Consequently, there is still a gap between pilots' intent to stay 20 miles away from hazardous storms and their ability to estimate distances accurately.

We also found that, on average, pilots underestimate the "out-the-window" visibility for a range of simulated visibilities. In Simulation 2, only 17% of the pilots could correctly report their lowest visibility encounter during flight. Furthermore, a majority of these correct reports (76%) were from pilots who entered ~ one nmi visibility zones (i.e., IMC). This reveals a gap in pilot situational assessments and undermines a high WSA during flights in deteriorating visibility. The failure to assess current visibility conditions increases the odds of VFR-into-IMC flights.

Page Intentionally Left Blank

1. INTRODUCTION

This study was part of a larger project to assess the effect of cockpit weather presentations on General Aviation (GA) pilot performance and decision-making. In this study, we addressed the use of cockpit weather displays to reduce Visual Flight Rules (VFR) flights into Instrument Meteorological Conditions (IMC), a situation in which GA pilots intentionally or inadvertently navigate into areas where they can no longer fly solely by means of visual references (Batt & O'Hare, 2005; Wilson & Sloan, 2003).

While there are many factors contributing to VFR-into-IMC, Goh and Wiegmann (2001) specifically mention situation assessment, risk perception, motivation, and decision framing. In this study, we focused on the potential importance of situation assessment. In terms of weather, situation assessment relates to several factors, such as a pilots' cognitive framing of the current weather situation, weather conditions in the near future, the effect of current and future weather conditions on the aircraft, the likelihood of a safe flight, and available alternatives for turning around or deviating to alternate airports. A correct weather situation assessment implies that pilots first integrate all available information and then make informed decisions on whether or not to continue the flight. However, there are circumstances in which the weather situation assessment can go wrong. For example, pilots might not detect deteriorating visibility solely from the "out-the-window" view, leading to VFR-into-IMC (Wiggins & O'Hare, 2003). Similarly, pilots may fail to detect, integrate, or use additional information like Automated Weather Observing Systems (AWOS) reports or available weather displays.

There are many potential weather hazards that can lead to VFR-into-IMC. Common adverse conditions are clouds, fog, or hazardous precipitation (Hunter, Martinussen, Wiggins & O'Hare, 2011; Wilson & Sloan, 2003). On cockpit weather displays, these conditions are frequently indicated by METAR symbols for ceiling and visibility, and Next-Generation Radar (NEXRAD) depictions for areas of precipitation. This information could aid a pilot during situational assessments and contribute to enhanced weather situational awareness during flight. However, previous research has uncovered instances in which it seems like pilots are failing to detect, integrate, and use graphical METAR and NEXRAD information to enhance situational assessments and flight decisions.

A study by Johnson, Wiegmann, and Wickens (2006) used METAR symbols that indicated ceiling, wind, visibility, and flight category information on moving map displays. They found a very modest effect of the METAR information with only 7% of pilots using the information to avoid deteriorating weather. Similarly, Coyne, Baldwin, and Latorella (2005) studied METAR information using color-coded symbols with visibility and ceiling information. They found that the METAR information negatively influenced some pilot judgments of ceiling and visibility, leading to an overestimation of the out-the-window weather conditions.

According to the situation assessment hypothesis (Goh & Wiegmann, 2001), pilots make weather assessment errors (i.e., misdiagnosis) that lead to failures in accurately assessing hazardous conditions. This could explain the unsuccessful integration of displayed METAR information with perceived ceiling and visibility conditions. It could also explain the difficulty pilots have in estimating the out-the-window ceiling and visibility conditions. For example, Coyne, Baldwin, and Latorella (2008) found that pilots have great difficulty in accurately estimating out-the-window weather conditions from short (5 sec) simulated weather scenarios using four levels of ceiling (i.e., 400, 900, 2900, and 4500 ft.) and four levels visibility (i.e., 2, 3, 5, and 10 miles).

They found that pilot estimation errors are often in the direction of overestimating weather conditions. Furthermore, pilots have difficulty making independent estimates of ceiling and visibility conditions. When the simulated ceilings were low, pilots underestimated the out-the-window visibility. When the ceilings were high, pilots tended to overestimate the out-the-window visibility. Another report by Goh and Wiegmann (2001) also demonstrates pilots' difficulty interpreting out-the-window visibility. They found that, on average, pilots who decided to divert their flight had fewer visibility estimation errors (mean error of 0 miles) than pilots who decided to continue their flight (mean error of 1.4 miles). Furthermore, pilots who continued their flight also tended to overestimate the out-the-window visibility. A less clear distinction between pilot overestimation and underestimation with regard to visibility was reported by Wiegmann, Goh and O'Hare (2002). Using a GA cross-country weather scenario, they found that 35.3% of the pilots were accurate in their estimates of out-the-window visibility; 26.5% overestimated the visibility; and 38.2% underestimated the visibility. These results clearly indicate a gap in pilots' ability to accurately assess weather conditions. This situation assessment gap might also explain why pilots get too close to hazardous thunderstorms despite the use of modern NEXRAD displays.

Beringer and Ball (2004) investigated the use of NEXRAD displays and found that some pilots attempted to navigate between precipitation cells, which is a very unsafe maneuver for GA pilots. Furthermore, 53% of the study pilots failed to keep safe distances from hazardous precipitation cells. Burgess and Thomas (2004) reported a similar result in a study on the effect of improved cockpit weather displays on GA pilot decision-making and weather avoidance. The results showed no meaningful difference in distance-to-weather between a control group without a weather display and two groups using either a NEXRAD product or a National Convective Weather Forecast product. All three groups failed to keep safe distances to hazardous storm cells. Wu, Gooding, Shelley, Duong, and Johnson (2012) reported the same failure to keep safe distances from storms in a study of pilot decision-making during convective weather.

In 2014 the Weather Technology in the Cockpit (WTIC) program funded and successfully completed WTIC II, which explored the effects of variations in cockpit weather symbology on GA pilot performance and decision-making (Ahlstrom & Suss, 2014). Results originating from WTIC II highlighted a performance gap in which variations in METAR symbols affected pilot perception and decision-making. During the simulation, researchers introduced three METAR symbol changes that indicated reduced ceiling and visibility conditions at airports (i.e., a change from visual meteorological conditions (VMC) to instrument meteorological conditions (IMC). The outcome showed that pilots varied considerably in their detection of METAR symbol changes during flight. Among the three pilot groups, detection performance ranged from 25% to 62% depending on the METAR symbol color (blue/yellow or white/red) and shape (triangle or circle). This variability in the detection of METAR changes led to a variability in pilot decisions to continue the flight using VFR, contact air traffic control and inquire about weather updates, and request an instrument flight rules (IFR) flight plan. These results indicate a gap in pilots' ability to extract and use METAR information to enhance their situational assessment and decision-making.

In 2015, the WTIC program funded and successfully completed WTIC III, which explored the effects of portable weather applications and weather notifications on GA pilot performance and decision-making (Ahlstrom, Caddigan, Schulz, Ohneiser, Bastholm & Dworsky, 2015a, 2015b). During one of the simulation flights, pilots were randomly assigned either to an experimental

group (using a portable weather application that included a NEXRAD display) or to a control group (no weather application) and flew a simulated VFR flight while avoiding hazardous weather. The results showed that pilots with the portable weather application had credibly larger route deviations and credibly greater distances to hazardous weather ($\geq$ 30 dBZ cells) than the control group. Nevertheless, both groups flew too close to hazardous weather cells with a closest point of approach to weather of 9.65 nmi and 6.93 nmi for the experimental and control groups, respectively.

During a second WTIC III simulation, pilots were randomly assigned either to an experimental group (equipped with a vibrating bracelet for METAR changes) or to a control group (without a vibrating bracelet). At two times during the scenario (12 min and 24 min), the METAR symbols at the destination airport and one alternate airport changed colors from blue (VFR) to yellow (IFR). During flight, both groups had access to a cockpit weather display with weather information covering the Flight Information Services-Broadcast basic products (graphical and text-based information). This included METAR and NEXRAD information. The simulation outcome showed that the state-change notifications (vibrating bracelet) alerted pilots in the experimental group about updated METAR information, with the result that these pilots had a credibly higher count of METAR display interactions than the control group (i.e., pilots were clicking on the METAR symbols to display the METAR text information). However, there was no credible difference between pilot groups with regard to the number of pilots or the scenario time for decisions to divert to an alternate airport. This was surprising since the experimental group received a METAR state-change notification at 12 minutes and another notification at 24 minutes into the scenario, but the average decision time for deviations did not occur until 28.8 minutes (control) and 32 minutes (experimental) into the scenario. It is possible that when pilots received a notification from the vibrating bracelet, they directed their visual attention to the weather display but failed to detect which METAR symbol had changed. This would explain the increased number of display interactions and the failure to use METAR information for early deviation requests. Furthermore, similar to the first WTIC III simulation flight, there was no efficient use of the NEXRAD information during weather avoidance. The average closest point of approach to hazardous weather cells was 5.29 nmi for the control group and 4.89 nmi for the experimental group, approximately four times closer than the recommended distance to storms. These results indicate a gap in pilots' ability to extract, integrate, and use METAR and NEXRAD information for accurate situational assessments, decision-making, and safety of flight.

In summation, previous research and the WTIC II–III simulation results identified several gaps in the ability of pilots to extract and use METAR and NEXRAD information to enhance situation assessments and decision-making. During single-pilot operations, METAR symbol information and METAR symbol changes were difficult to detect. Therefore, pilots failed to extract and use important ceiling and visibility information. During weather avoidance flights, NEXRAD displays did not seem to provide pilots with an accurate situation assessment of storm location or their relation to the storm.

This study addresses these performance gaps by evaluating the effect of 1) symbol enhancements and 2) active display reminders for visibility and precipitation thresholds on pilot visual performance and flying behavior.

The study consisted of two between-subjects flight simulations. In Simulation1, participants were randomly allocated to one of two enhanced weather presentations. The flight scenario was specifically designed to assess pilots' sensitivity to changes in METAR symbols during flight. In Simulation 2, we assessed the effect on pilot flying behavior and decision-making through manipulations of visibility along the route of flight. For Simulation 2, participants were randomly allocated to one of two groups, one of which used a portable weather application that provides an active reminder (AR) of forecasted visibility conditions and the distance from the aircraft to hazardous precipitation.

1.1 Purpose

The overarching goal of the study was to evaluate how weather display information affects pilot performance and decision-making during single-pilot operations. The particular objectives of this study were:

1. to evaluate the impact on pilot behavior and decision making from METAR display information during flights in deteriorating visibility (Simulation 1), and

2. to evaluate the impact on pilot behavior and decision-making from the use of an AR with storm and visibility information during flights in deteriorating visibility (Simulation 2).

2. SIMULATION 1

In Simulation 1, we evaluated pilot perception of METAR symbol changes during flight in deteriorating weather and assessed how this information affects pilot flight behavior and decision-making.

In a study of METAR display information, Ahlstrom and Suss (2014) used three display alternatives (called D1, D2, and D3) for METAR information that used different symbols and colors. The result showed that pilots had difficulty detecting METAR color changes that corresponded to a transition from VFR to IFR conditions. The overall METAR detection performance for the D1 pilot group was only 25% whereas it was 62% for the D3 group. Besides a difference in the METAR colors (i.e., yellow/blue and white/red), the D1 and D3 representations also differed in their symbol shapes. The D1 symbology used triangular METAR symbols whereas the D3 symbology used circular METAR symbols. In addition to the low detectability, the results also showed that pilots were credibly better at detecting color changes to circular METAR symbols (D3) than triangular METAR symbols (D1) during flight.

Therefore, in Simulation 1, we assessed enhanced (i.e., increased METAR color salience) D1 and D3 symbologies on pilot detection performance and whether there are any differences in detection between enhanced METAR triangles (D1) and enhanced METAR circles (D3).

Our hypothesis was that the enhanced D1-D3 METAR symbols would yield a higher detection rate than the Ahlstrom and Suss (2014) result and that there would be equal detection performance for METAR triangles (D1) and METAR circles (D3) during flight.

2.1 Method

2.1.1 Participants

In many studies, researchers recruit participants according to a convenience sampling technique (Etikan, Musa, & Alkassim, 2015), which means researchers are recruiting participants that are

easy to access and available in great numbers. One of the rationales for using this nonrandom sampling technique is that the target population is so large that it would be impossible to study each participant. Another rationale is that researchers assume that participants from the population in question are homogeneous.

However, convenience samples are rarely representative of the target population. The target population for this study is GA private pilots. As a population, the members of this group are not homogeneous in important variables like age and flight experience. Figure 1 shows the predicted number of active private GA pilots in the United States (FAA, 2016).

2015 Private pilots (total 186,786 pilots)
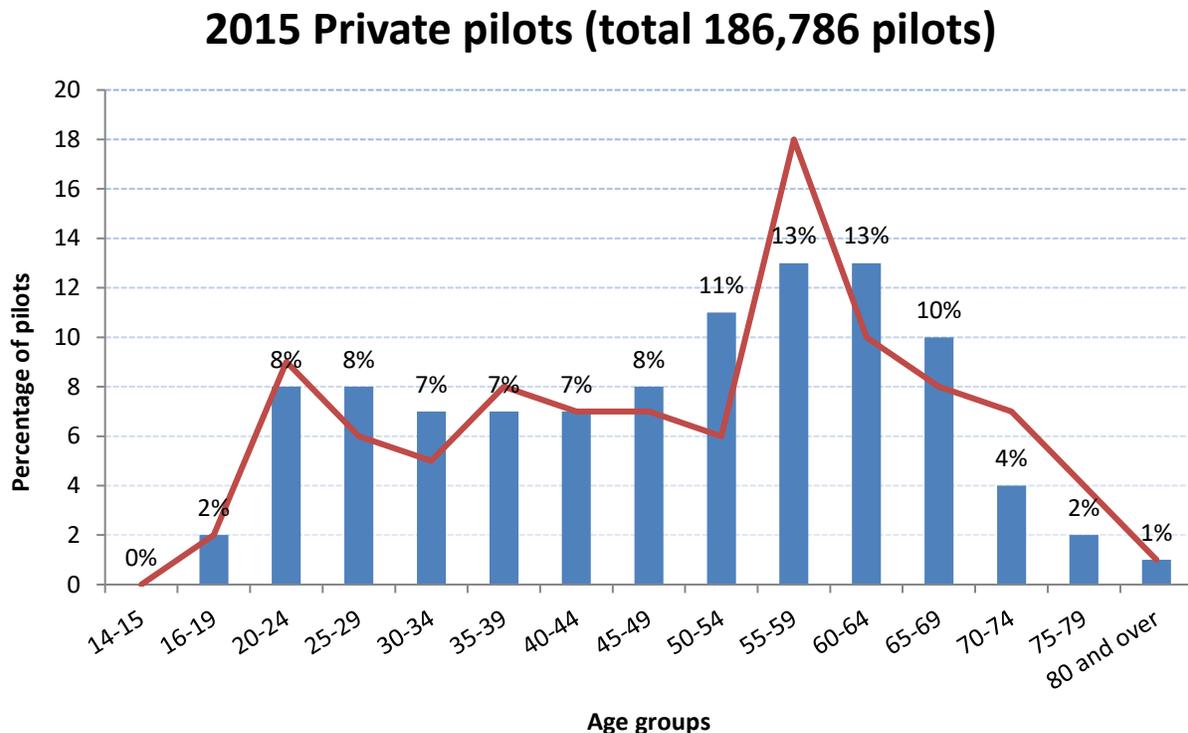
Figure 1. Estimated active private pilot certificates by age group as of December 31, 2015 (data from FAA, 2016). The red line represents the sample of private pilots used in this study.

In the FAA data, the total number of estimated active private pilots is divided into five-year age groups starting at 14–15 years old, and ending at 80 years and older. As we can see in Figure 1, this yields a distinctive distribution of pilot ages. From previous studies, we know that it is nearly impossible to achieve this sampling distribution from convenience sampling; the sampling distribution will usually end up being skewed with an overrepresentation of older pilots or an overrepresentation of younger pilots. This is problematic because previous research of cognitive proficiency and flight performance has shown the presence of pilot age-group differences for perceptual-motor skills, memory, attention, and problem solving (Hardy & Parasuraman, 1997). Therefore, a non-representative sample is not desirable if we want to make inferences about the target population.

To alleviate some of these problems we used stratified random sampling (Barnabas & Sunday, 2014) when recruiting participants. First, we conducted a power analysis on the precision (Highest Density Intervals [HDI] of differences within -0.2 to 0.2) by which pilots can detect the

presence of METAR symbol changes during flight (based on data and a power analysis framework from Ahlstrom and Suss, 2014). We wanted to estimate the number of participants needed to reach a power of 80% for achieving our goal. The outcome revealed an 80% power in achieving our goal using 90 participants. Second, we used the N = 90 to compute sample percentages for each of the age strata shown in the Figure 1 histogram. Third, we randomly sampled the number of private GA pilots required for each age strata.

Because of technical problems during the first week of simulations, we had to replace pilots for some of the age strata. This resulted 93 private GA pilots for Simulation 1 and 95 private GA pilots for Simulation 2. Our resulting age distribution sample is superimposed as a red line on the FAA data in Figure 1. Although not perfect, our pilot age group sample is very close to the estimated age groups for private pilots in the National Airspace System.

Because Simulation 1 used a between-subject design, each pilot was randomly allocated to one of two simulation groups (e.g., control and enhanced). We present the participant background information as group summaries in Table 1.

Table 1. Descriptive Characteristics of Simulation 1 Participants

| | | Age (years) | | Flight hours accrued | | | | | |
| | | | | Total | | Instrument | | Instrument- last 6 mo. | |
Group	n	Median	Range	Median	Range	Median	Range	Median	Range
Control	47	48	18-82	325	78-5000	9	0-1000	0.5	0-55
Enhanced	46	53.5	20-76	300	65-4500	15	0-1000	2.5	0.89.6

2.1.2 Informed Consent Statement

Each participant read and signed an informed consent statement before starting the study (see Appendix A). Informed consent statements describe the study, the foreseeable risks, and the rights and responsibilities of the participants, including that their participation in the study is voluntary. All the information that the participant provides, including personally identifiable information, will be protected from release except as may be required by statute. Signing the form indicates that the participant understands his or her rights as a participant in the study and their consent to participate.

2.1.3 Research Personnel

The research team (see Appendix B) consisted of human factors researchers, Air Traffic Control (ATC) Subject Matter Experts (SMEs), pilot SMEs, computer scientists, and aerospace engineers. The team prepared scenarios, briefings, and collected data during the simulations.

2.1.4 Facilities

We conducted the study at the FAA William J. Hughes Technical Center Cockpit Simulation Facility.

2.1.5 Aircraft Simulator

For Simulation 1, we used a generic GA aircraft simulator (fixed base) with a 180° out-the-window view provided by six vertical 52-inch Samsung Liquid Crystal Display monitors (see Figure 2), configured to simulate a Cessna 172 single-engine aircraft with the G1000 type GA glass cockpit control display. We used X-Plane 10 for both the flight dynamics and the out-the-window view. The simulator was equipped with a stand-alone portable weather display running

on a Windows Surface Pro 3 and a voice communication system that provided a link between the pilots and ATC through a Push-To-Talk (PTT) capability.

Figure 2. Left: an exterior view of the generic GA aircraft fuselage. Right: the cockpit out-the-window view, the G1000 type GA glass cockpit control display, the instrument (radio) stack, and the stand-alone weather display running on a Windows Surface Pro 3.

2.1.6 Video and Sound Recordings

During the study, the cockpit simulator was equipped for video recording (H.264 format) and sound recording and playback. To capture pilot behavior we used three cockpit-mounted cameras that provided a top view (dome camera, 360°), a front view (bullet camera), and a side view (fisheye camera). These three camera views captured the entire cockpit environment and the videos were therefore suitable for a behavioral analysis of pilot actions during flight. In addition to the camera views, we also captured and recorded the display of the G1000 type GA glass cockpit control display and the auxiliary weather display (Microsoft® Surface Pro 3).

The cockpit simulator was on a separate Local Area Network and we used Real Time Streaming Protocol to capture live Internet Protocol camera streams. We used the iSpy surveillance software for recording and for video playback at the researcher control station (see Figure 3). The iSpy software synchronized the recordings of video and sound, and displayed the five individual video streams from the cockpit simulator. The cockpit sound system captured voice recordings from pilots and ATC, and allowed playback of prerecorded AWOS messages.

Figure 3. The researcher control stations.

2.1.7 Radio Simulation Software

We used the PLEXYS software to simulate radio PTT communication between the pilot and ATC. PLEXYS also managed the playback of prerecorded live ATC/pilot sector communication. In addition, PLEXYS managed the synchronization and playback of AWOS messages during the simulation flights.

2.1.8 Data Handling Procedure

We assigned a coded identifier to each participant pilot. The identifier did not appear on the Informed Consent Statement because that is identified by the participant's signature. We tagged all other data collection forms, computer files, electronic recordings, and storage media containing participant information only with the coded identifier, not the name or personal identifying information of the participants. We retained original documents, recordings, and files as collected. All data-editing, cleanup, and analysis was performed on copies traceable to the original sources.

2.1.9 Data Analysis

The data from this study were analyzed using Bayesian estimation as used in Ahlstrom and Suss (2014) and Ahlstrom et al. (2015a; 2015b). During the analysis we used JAGS ("Just Another Gibbs Sampler": Plummer, 2003, 2011) that we called from R (R Development Core Team, 2011) via the package RJAGS. All software for the analysis and figure generation was adapted program code from Kruschke (2014).

The Bayesian analysis generates a posterior distribution, which is a distribution of credible parameter values. We can use this large distribution of representative parameter values to evaluate certain parameters, or to compare differences between parameters. Here, we used a separate decision rule to convert our posterior distributions to a specific conclusion about a parameter value. When plotting the posterior distribution, we included a black horizontal bar that represented the 95% HDI. The HDI had a higher probability density compared to values that fall outside the HDI. When we compared conditions (i.e., perform contrasts), we computed differences at each step in the Markov Chain Monte Carlo (MCMC) chain and present the result

in a histogram with the HDI. These histograms showed both credible differences and the uncertainty of the outcome. If the value 0 (implying zero difference) was not located within a 95% HDI, we said that the difference is credible. If the 95% HDI included the value 0, the difference was not credible as it meant that a difference of zero was a possible outcome.

For effect size analyses, we also used a Region of Practical Equivalence (ROPE). The ROPE contains values that, for all practical purposes, are the same as a null effect (i.e., no meaningful difference). If the 95% HDI fell completely within the ROPE margins for an effect size, we could declare the presence of a null effect, and unlike traditional analyses, we could accept the null outcome. If, however, the entire ROPE fell outside the 95% HDI, we could reject the presence of a null effect.

The various data sets from this study are not on the same measurement scale. For example, some data sets are on a continuous metric scale whereas others are on a discrete non-continuous scale. Because of this, we are using different Bayesian models that we describe for each analysis in the result section.

To derive the posterior distributions we used 200,000 samples (regardless of model). For all analyses, we used priors that were vague and noncommittal on the scale of the data.

2.1.10 Flight Plan

The VFR flight was planned to depart from Allentown, Pennsylvania (KABE) under flight-following, and to reach the destination airport at Martinsburg, West Virginia (KMRB). The flight plan followed a Very High Frequency (VHF) Omni Directional Radio Range (VOR) route (i.e., VOR-to-VOR) from KABE to East Texas VOR (ETX), Lancaster VOR (LRP), VINNY intersection, Westminster VOR (EMI), Martinsburg VOR (MRB), and then to KMRB (see Figure 4). As the goal of the study was to investigate the effect of the METAR symbology on pilot detection and decision-making during at-altitude flight, we started the aircraft at a cruising altitude of 8500 feet in the area of KABE headed toward ETX (heading 223) at a speed of 123 knots; we used the local time 2:00 PM in the month of June.

At any time during the scenario, IFR-rated pilots could request an IFR flight plan.
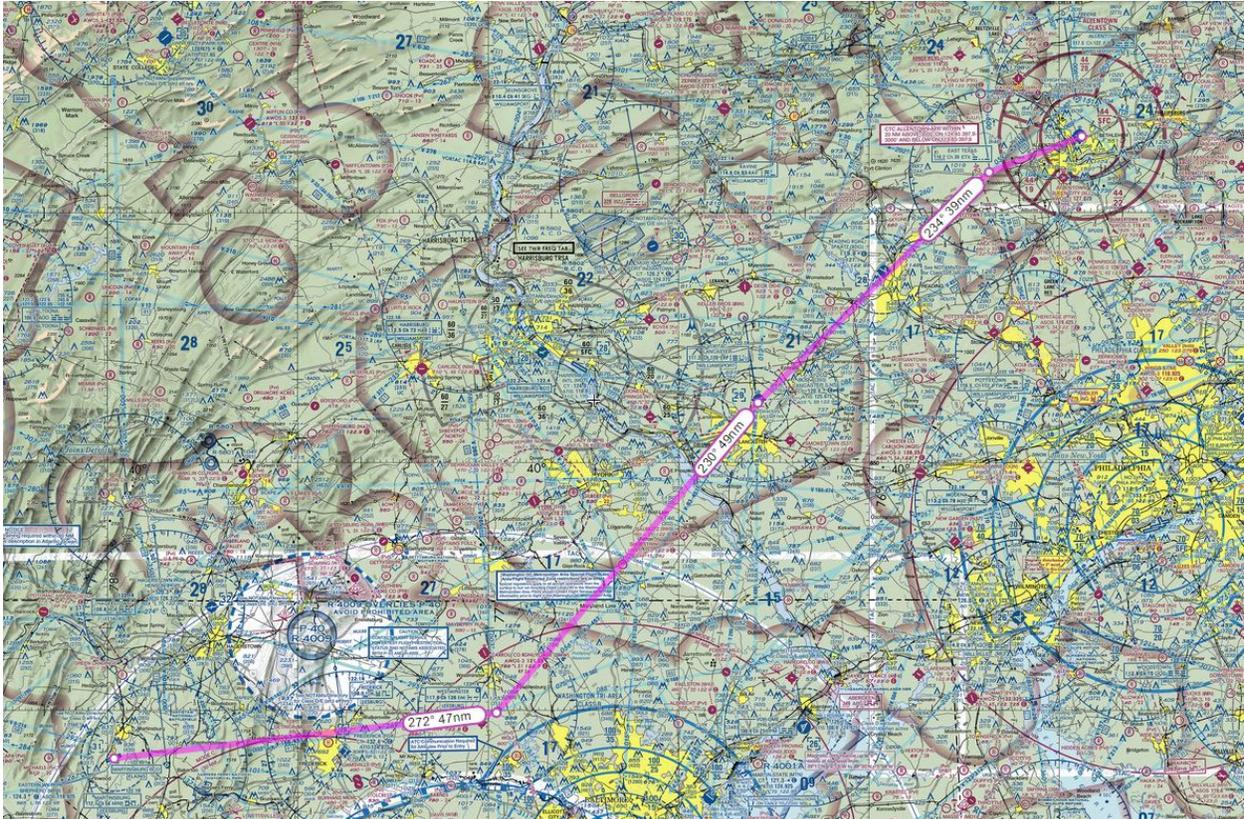
Figure 4. Scenario route from Allentown (KABE, top right) to Martinsburg airport (KMRB, bottom left).

2.1.11 METAR Symbol Changes

For this simulation, we were interested in the effect of enhanced symbology on pilot VFR-to-IFR change detection. At the start of the scenario, all METAR symbols along the route indicated VFR conditions. To assess pilot detection of changes in METAR status, selected METARs in the region of the planned route were programmed to change from VFR to IFR at three time points during the 35-minute flight. The first METAR change occurred at 10 minutes into the scenario, the second at 19 minutes, and the third and last METAR change occurred at 30 minutes (see Figure 5). Table 2 describes the timing of the changes, and the METAR(s) that changed at each point. After a METAR changed from VFR to IFR, it remained IFR for the remainder of the scenario. In addition to these METAR changes, pilots had the ability to tune into six AWOS stations to receive real-time weather information (i.e., MRB, HGS, BWI, ESN, IAD, and DCA).
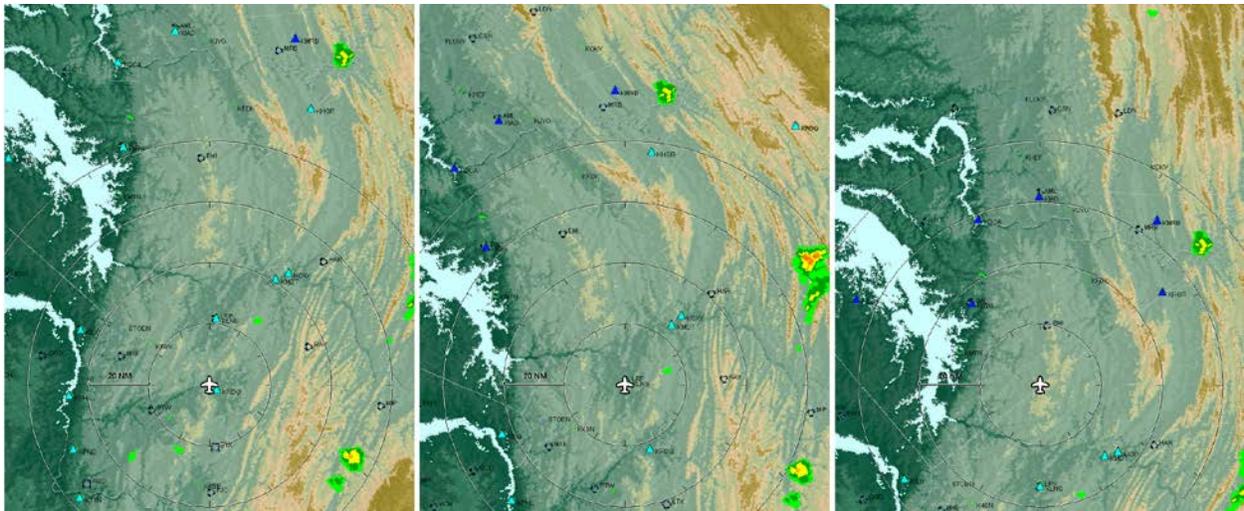
Figure 5. Illustration of the METAR changes during flight. VFR METARs are shown as light blue triangles and IFR METARs as dark blue triangles. Left: At 10 minutes into the scenario, the METAR at the destination airport changed to IFR. Middle: At 19 minutes into the scenario the METARs at KBWI, KDCA, KESN, KIAD, and KJST changed to IFR. Right: At 30 minutes into the scenario, the METAR at KHGS changed to IFR.

Table 2. METAR Change Times and Locations

			METAR changes from VFR to IFR	
Change #	Scenario time (minutes)	Number of METARS changed	METAR code	Location
1	10	1	KMRB	Martinsburg, WV
2	19	5	KBWI	Baltimore/Washington International, MD
			KDCA	Washington National, DC
			KESN	Easton/Newman, MD
			KIAD	Washington, DC/Dulles, VA
			KJST	Johnstown, PA
3	30	1	KHGS	Hagerstown, MD

2.1.12 Enhanced METAR Symbology

METAR symbols can be broken down into four weather parameters based on the flight category (see Table 3). For this study, we only used two METAR symbols (labeled D1 and D3) that indicated VFR or IFR (Ahlstrom and Suss, 2014). In the Ahlstrom and Suss study, the D1 METAR symbology displayed light blue triangles for VFR and yellow triangles for IFR (see Figure 6, top left), and the D3 METAR symbology displayed white circles for VFR and red circles for IFR (see Figure 6, top right).

Table 3. Flight Categories (FAA & NOAA, 2010)

Category	Ceiling		Visibility
Low Instrument Flight Rules (LIFR)	< 500 feet AGL	**And/Or**	< 1 mile
Instrument Flight Rules (IFR)	500 to 1,000 feet AGL	**And/Or**	1 mile to 3 miles
Marginal Visual Flight Rules (MVFR)	1,000 to 3,000 feet AGL	**And/Or**	3 to 5 miles
Visual Flight Rules (VFR)	> 3,000 feet AGL	**And/Or**	> 5 miles
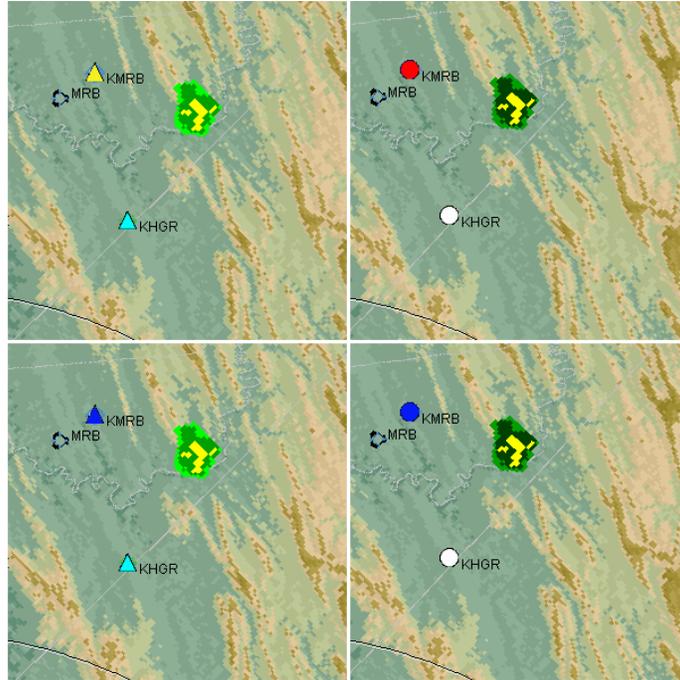
Figure 6. Original D1 (top left) and D3 (top right) METAR colors used by Ahlstrom and Suss (2014). Enhanced D1 (bottom left) and D3 (bottom right) versions used in this simulation.

Because our goal was to enhance the METAR symbol colors, we performed an analysis of the original symbols used by Ahlstrom and Suss. First, using an algorithm by Achanta, Hemami, Estrada, and Süsstrunk (2009), we performed a frequency-tuned salience analysis of the D1 and D3 colors. The output of this algorithm was a high-resolution salience map in which bright objects have the highest salience and dark objects the lowest salience. We assessed the salience by measuring the METAR symbol intensity on the salience map (index from 0–255). Our analysis revealed that the original D1 VFR symbols had an approximate mean salience of 61 (*SD* = 1.3) and the D1 IFR symbols had an approximate mean salience of 112 (*SD* = 1.0). Therefore, during a D1 METAR symbol change from VFR to IFR, the average symbol salience change was approximately 51. Analyzing the D3 symbol colors, we found that the VFR symbols had an approximate mean salience of 30 (*SD* = 0.4) and the IFR symbols had a mean salience of 151 (*SD* = 0.5). Therefore, during a D3 VFR-to-IFR change, the approximate mean salience change was 121. In the Ahlstrom and Suss study, the higher salience difference for the D3 VFR-to-IFR change compared to the D1 VFR-to-IFR change did affect pilot performance. The average

METAR change detection rate for D1 METARs during flight was only 25% whereas it was 62% for D3 METAR symbols.

To enhance the METAR symbols in this study we manipulated the original D1 and D3 IFR colors as shown in the bottom row of Figure 6. For the D1 METAR symbols (bottom left), we increased the salience of the original IFR METAR symbols (yellow triangles) from a mean salience of 112 to a mean salience of 251 ($SD = 0.75$ dark blue triangles). This means that for an enhanced D1 METAR change from VFR to IFR the average METAR salience change was approximately 190.

For the D3 METAR symbols (see Figure 6, bottom right), we increased the salience of the IFR METAR symbols (red circles) from a mean salience of 151 to a mean enhanced salience of 252 ($SD = 0.52$, dark blue circles). This means that for an enhanced D3 METAR change from VFR to IFR the average METAR salience change was approximately 222.

Because our enhanced symbols provide a larger change in the METAR salience during VFR-to-IFR changes, we predicted an increased pilot detection performance in the simulation compared to the Ahlstrom and Suss result. Furthermore, because the changes from VFR-to-IFR produce similar salience differences for the enhanced D1 and D3 symbologies, we predicted that pilot performance would be similar for METAR triangles (D1) and METAR circles (D3).

2.1.13 Weather Scenario – Reductions in Visibility

To assess pilot Weather Situation Awareness (WSA) during flight, we created an out-the-window scenario in which the visibility decreased as pilots flew toward the destination airport. As shown in Figure 7, the pilots began the scenario in the 50 nmi visibility zone. However, the visibility conditions along the route were deteriorating with the visibility progressively reduced to zones with 30 nmi, 20 nmi, 10 nmi, 8 nmi, 4 nmi, and finally 2 nmi at the destination airport (see Figure 8). None of the pilots reached the destination airport due to the 35 min simulation time.

We created the visibility zones by feeding selected METAR strings to X-Plane (see Table 4).

Table 4. Airport METAR Strings

Airport	Identifier	METAR String
Lehigh Valley International Airport	KABE	102100Z 26306KT 50SM OVC100 A2992
Reading Regional Airport/Carl A Spaatz Field	KRDG	102100Z 26306KT 30SM OVC100 A2992
Lancaster Airport	KLNS	102100Z 26306KT 20SM OVC100 A2992
York Airport	KTHV	102100Z 26306KT 10SM OVC100 A2992
Carroll County Regional Airport	KDMW	102100Z 26306KT 08SM OVC100 A2992
Frederick Municipal Airport	KFDK	102100Z 26306KT 04SM OVC100 A2992
Eastern WV Regional Airport/Shepherd Field (Martinsburg)	KMRB	102100Z 26306KT 02SM OVC100 A2992

X-Plane read these METAR strings and created the out-the-window visibility during the scenario start up. None of the METAR strings changed; they were static throughout the scenario. The result was a smooth transition in visibilities as pilots flew across the simulated visibility zones. The out-the-window view for all scenario runs was identical.

Figure 7. View from the cockpit at minute two into the PA scenario.
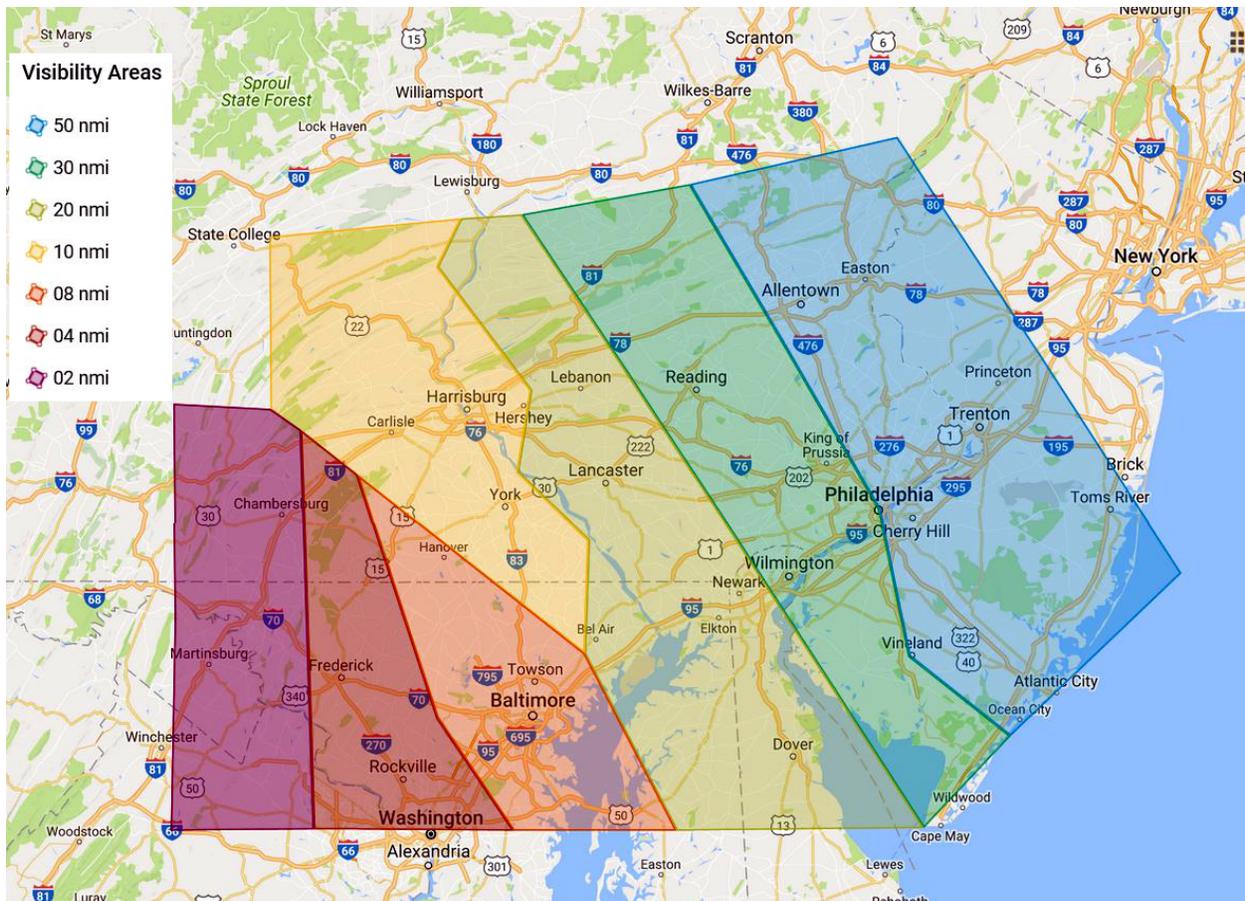
Figure 8. Areas of simulated visibility along the route from Allentown to Martinsburg Airport.

2.1.14 Simulation Procedure

After reading and signing the Informed Consent Statement (see Appendix A) participants completed a Biographical Questionnaire (see Appendix C), a Mobile Device Proficiency Questionnaire (see Appendix D; Roque and Boot, 2016), and a Weather Knowledge Questionnaire (see Appendix E). After finishing the questionnaires, researchers briefed participants on the schedule of events and explained the general purpose and procedures of the simulation. After this brief introduction, participants completed a self-paced pre-flight briefing (presented as a Microsoft® PowerPoint® slideshow). The flight briefing contained general information about the purpose of the study. It also contained a detailed overview of the cockpit simulator, the G1000 display, the auxiliary weather display, the instrument stack, the Horizontal Situation Indicator (HSI), the AWOS radio frequencies, and the flight plan.

Following the self-paced flight briefing, participants were given unlimited time to read the flight reference materials, which included a sectional map, the flight plan, relevant ATC and AWOS radio frequencies, and a printed weather briefing.

After studying the flight reference material a researcher escorted the participant to the cockpit simulator for a 15-minute practice flight, during which the researcher familiarized the participant with the simulator's (a) aircraft controls, (b) weather presentation, (c) radio, (d) HSI, (e) AWOS

radio frequencies, and (f) navigation equipment. When the pilot was ready to begin the simulation, the researchers started the 35-minute data collection flight.

At the predetermined freeze points (i.e., 11, 20, 35 minutes) we froze the simulator (i.e., paused) and presented the participant with a probe question window that covered the G1000 display (see Appendix F for the specific questions at each freeze point). Simultaneously with the simulator freeze, we also presented a blue neutral background on the auxiliary weather display. The covering of the G1000 display and the neutral weather display background meant that pilots were unable to use any of this information when answering the probe questions. The participant read each question aloud and provided an answer. As soon as participants had answered all the questions, we resumed the simulation. After participants answered the probe question at the 35-minute freeze point, researchers terminated the simulation.

At the conclusion of the simulation flight, participants completed a post-scenario questionnaire (see Appendix G).

2.1.15 Independent Variable

We manipulated the independent variable -*Weather Symbology*- by presenting METAR information under two different symbology modes (D1 and D3). Half of the participants flew the scenario using the D1 presentation whereas the remaining participants used the D3 presentation.

2.1.16 Dependent Variables

During Simulation 1, we recorded key dependent variables to evaluate pilot detection of METAR changes during flight, pilot WSA, and how the METAR changes and reduced visibility affected pilot decision-making and flight behavior.

Table 5. Dependent Variable List

Number	Dependent variable	Description
1	System performance measures	Data from the cockpit simulator (e.g., altitude, heading, lat/long position).
2	Pilot/ATC communications	The content of pilot/ATC communications.
3	Weather situation awareness	- SAGAT query of the number of times pilots detected a METAR color change (at 11 min, 20 min, and 35 min into the scenario).
		- The use of Automated Weather Observing System (AWOS) stations.
		- Pilot visibility reports (at 11 min, 20 min, and 35 min into the scenario).
4	Decision-making	- Pilot decision to continue flight, turn around, or to land at an alternate airport.

The dependent variables captured the following categories: Altitude Profiles (simulator data), Communication (pilot/ATC PTT), Weather Situation Awareness (detection of METAR changes, pilot visibility estimates, and the use of AWOS stations), and Decision-Making (e.g., whether the

pilot continued flight, turned around, or deviated to an alternate airport). In Table 5, we provide a list of the dependent variables and a short description.

To ascertain whether pilots had detected the METAR changes, we used a modified version of the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1995). The SAGAT involved administering a series of targeted probe questions during brief, temporary freezes in simulated scenarios.

2.2 Results

2.2.1 ATC Communication

During the simulation, we automatically recorded all pilot/ATC PTT communications. To complement these PTT transmissions, researchers took notes that provided additional information about the scenario time, communication content, and other relevant observations concerning pilot questions, behavior, and decision-making.

Using the researcher notes, we summarized pilot communications into five important categories. The first category includes communications related to pilot questions about IFR procedures or pilot requests to file an IFR flight plan (for IFR-rated pilots). The second category included communications in which the pilot reported IMC conditions along the route of flight or at selected airports. The third category included communications about pilot altitude requests. The fourth category included pilot communications regarding decisions or questions about turning around or deviating to an alternate airport. Finally, the fifth category included communications about going off the frequency to tune in to one or more AWOS stations.

Figure 9 shows a summary of the total communication frequencies (i.e., count across all runs) for each of the five categories.
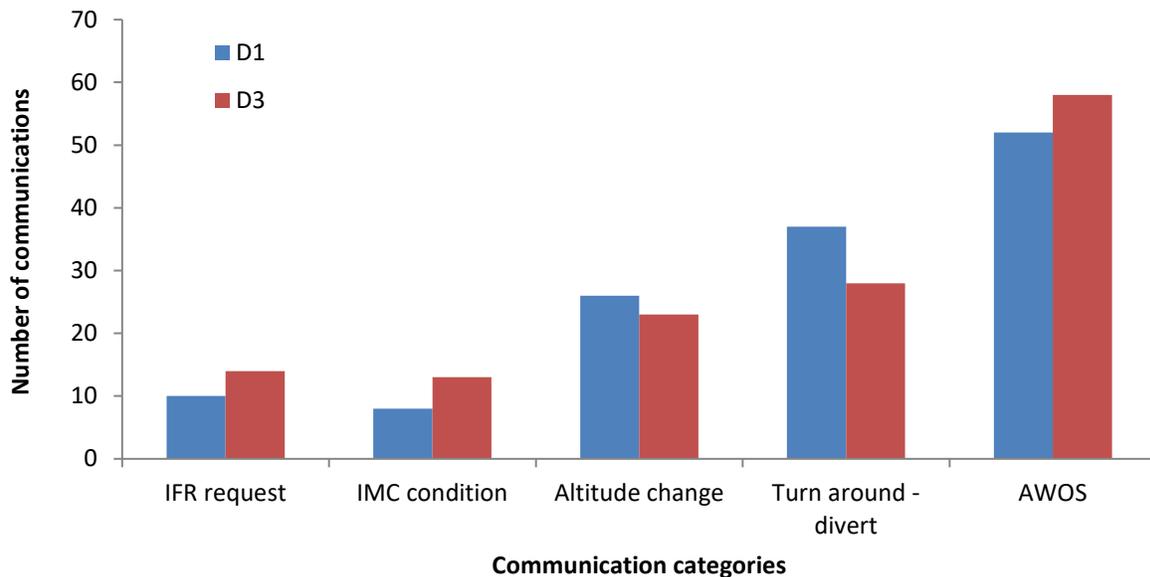
Figure 9. Summary of pilot communications for five communication categories.

As is shown in Figure 9, the category counts are very similar with the highest counts for AWOS communications, the lowest counts for IMC conditions and IFR requests, with the number of

altitude and turn around-divert communications in between. Most importantly, there are no large and important deviations in frequency counts between D1 and D3.

In summation, an analysis of the pilot/ATC communication contents reveals a similar communication count for the D1 and D3 groups. We find the highest counts for AWOS communications and the lowest counts for IMC conditions and IFR requests, with altitude and turn around-divert communications in between.

2.2.2 Weather Situation Awareness

In the following section, we present results from an assessment of pilot WSA. Pilot WSA is a complicated construct that, at the very least, involves pilot perception of time, airspace, route, current weather conditions, weather conditions in the near future, and available alternatives for turning around or deviating to alternate airports.

We used three different dependent variables to measure WSA: 1) the SAGAT query of the number of times pilot detected METAR changes during flight, 2) the use of AWOS information, and 3) pilot out-the-window visibility reports.

2.2.3 Detection of METAR Changes During Flight

During the simulation flights we froze the simulator at the 11-, 20-, and 35-minute marks and displayed the SAGAT questions (see Appendix F). The purpose of these questions was to assess whether the pilot had detected the METAR change(s). Table 6 shows the number of METAR detections for each SAGAT stop and METAR symbology. For the D1 METAR changes, there were $n = 46$ opportunities to detect METAR changes at each SAGAT stop for $N = 138$ opportunities. For D3, there were $n = 47$ opportunities at each SAGAT stop for $N = 141$ opportunities.

In Table 6, the METAR detection performance is shown to be very low for both the D1 and the D3 symbologies. The percentage of detections (out of all opportunities) is only 26% and 31% for the D1 and D3 symbologies, respectively.

Table 6. METAR Detection Data for D1 and D3 at the Three SAGAT Stops.

METAR symbology	SAGAT stops			
	11 min	20 min	35 min	Total
D1	5(46)	11(46)	20(46)	36(138)
D3	5(47)	15(47)	24(47)	44(141)

From the METAR detections in Table 6, we computed an overall detection score for each pilot based on the number of detections across the three METAR changes. To analyze this data we used a model from Kruschke (2011) for between-subjects analysis of dichotomous data. Applied to the current data, the model considers each data point as the outcome of each METAR change, in which the pilot gets zero for no detection and one for detection. Across the three METAR changes, each pilot gets a number (0–3) of METAR detections, with each pilot's propensity for detection being influenced by the individual's ability and the METAR symbology. The group level distribution of propensities is modeled as a beta distribution, with the groups' mean detection accuracy denoted by μ_c. Each pilot's spread around the group mean is denoted by the

parameter k_c, modelled with a broad gamma prior. In the current model implementation there is no higher-level model structure that expresses relations across the D1 and D3 conditions. Our prediction was that METAR detection accuracies are different for D1 and D3 so our main interest was in estimating differences in μ_c between the two symbology conditions.

Figure 10 shows the posterior distributions as a function of detection accuracy (μ_c) and precision (k_c). The detection accuracies are the same because the two distributions are superimposed, there is only a slightly lower precision (i.e., larger spread) for D1 (mean $k = 10.5$) compared to D3 (mean $k = 18.9$).
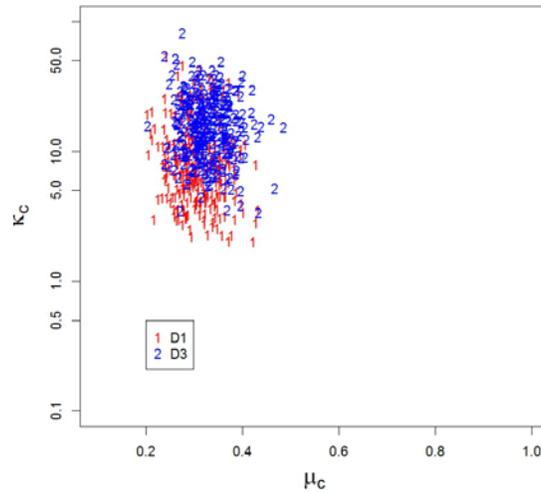
Figure 10. Mean posterior detection accuracies (μ_c) as a function of precision (k_c) for the D1 and D3 symbologies.

Figure 11 shows the posterior means and contrast for D1 and D3. The predicted performance is almost identical with a mean $\mu = 0.31$ for D1 and a mean $\mu = 0.327$ for D3. Consequently, the contrast shows the value 0 located in the middle of the 95% HDI and the difference distribution has a mean of -0.01.
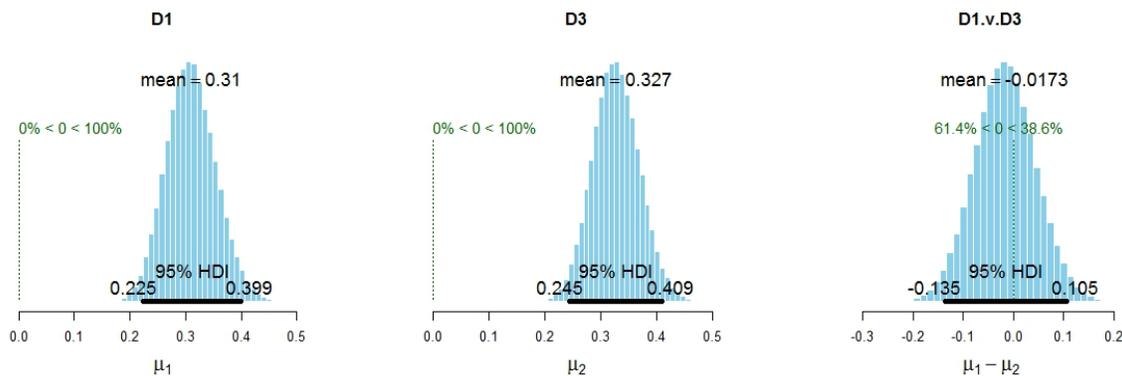
Figure 11. Mean posterior detection accuracies (left and middle) and contrast for the D1 and D3 symbologies.

The current result for the enhanced D1 and D3 METAR changes shows no difference in detection performance for triangles (D1) versus circles (D3). This is in contrast to the study by

19

Ahlstrom and Suss (2014), who found a credibly higher detection performance for METAR circles than METAR triangles. During their simulation flights, pilots using the D3 METAR symbology detected 62% of the METAR changes versus 25% for the pilots using the D1 METAR symbols. Therefore, the result shows that an enhanced METAR symbol salience eliminates the effect of METAR symbology (circles versus triangles).

To further our understanding of the efficiency of using METAR color changes (i.e., salience differences) to attract pilot attention we need to assess the performance outcome across several independent studies. If fact, there are several prior studies that have used the D1 and D3 METAR symbols for change detection. Besides this study and the study by Ahlstrom and Suss (2014), Ahlstrom et al. (2019) and Ahlstrom (2015) used the original D1 and D3 symbols in change-detection experiments. Therefore, with this study, there are data available from five independent studies. Two of these studies are METAR change detection during simulations flights and the other three are METAR change detection during part-task experiments. In all these studies, the D1 METAR symbols produced a lower salience difference for the VFR-to-IFR change than the D3 symbols.

To assess the outcome from all these studies, we used a hierarchical random-effects model for Bayesian meta-analysis by Kruschke and Liddell (2016). The model considers the outcome from two separate groups, like a control group and an experimental group, in which we have data from some result (e.g., correct detections) and the corresponding opportunities (e.g., total number of trials). For our purposes, we used the D1 and the D3 groups from the aforementioned studies. All these studies are similar and we therefore expected them to be mutually informative.

In the model, each individual study has its own parameters with higher-level distributions that estimate parameters across studies. The top-level distribution estimates the trend and dispersion across all independent studies. Contrary to the model used for the previous analysis, this hierarchical model implies that lower-level distributions inform the higher-level distribution. This means that the higher-level distribution constrains the parameters at lower levels to be mutually consistent. Therefore, the model imposes shrinkage toward the mode for extreme values. For a detailed description of the model assumptions, structure, and priors see Kruschke and Liddell (2016).

Figure 12 shows the effect of METAR salience (i.e., the difference between D1 and D3 METAR symbology) on the horizontal axis and the estimates of the five independent studies on the vertical axis. Study 1 is the data from the current flight simulations; Study 2 is the data from the Ahlstrom and Suss flight simulation; Study 3 is the data from the Ahlstrom et al. change-detection experiment; Study 4 is the data from the Ahlstrom and Suss change detection experiment; and Study 5 is the data from the Ahlstrom (2015) change detection experiment.
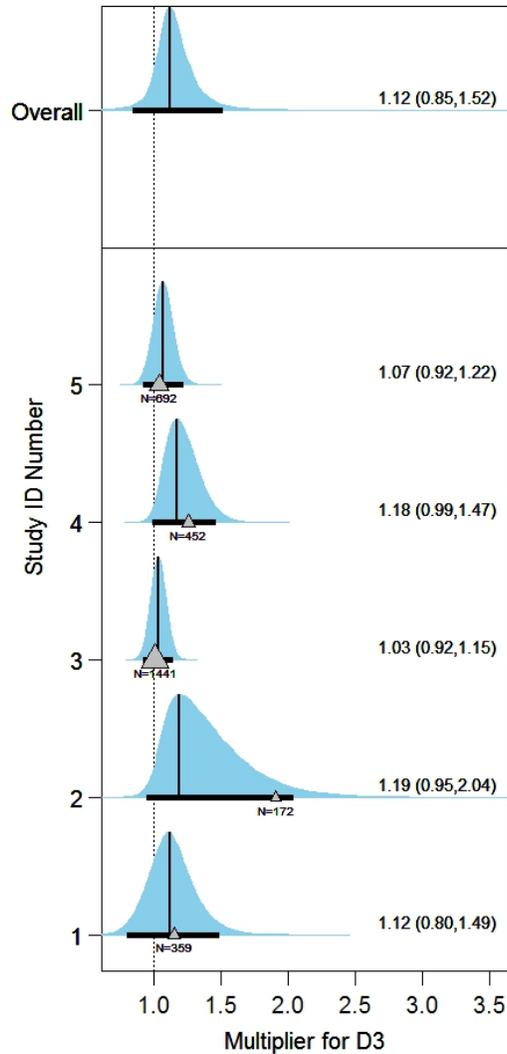
Figure 12. Meta-analysis of the effect of METAR salience on change detection. The overall estimate (top) is based on the five individual studies shown on the vertical axis.

At the top of Figure 12, we have the overall estimate of METAR detection performance, calculated across all the D1 and D3 trials from the five independent studies. As we can see, the estimated overall effect was 1.12, which tells us that the detection of D3 METAR symbols was 12% higher than the detection of D1 METAR symbols. We can also see that there was a lot of uncertainty in this overall estimate, with the 95% HDI ranging from 0.85 to 1.52. This was expected because our overall estimate was only based on data from five independent studies.

Each of the individual studies had its own sample value and sample size denoted by a triangle. The sample value measured the D3 detection rate as a proportion of the D1 detection rate. For most studies, the sample values and the predicted values were not very far from each other. In Study 1, for example, the sample value (at the triangle location) was only slightly larger than the estimated value denoted by the vertical line (i.e., the mode) in the posterior distribution. There is very little shrinkage for the Study 1 estimate. For Study 2, however, the sample value (triangle) was far from the estimated value (mode) of the posterior distribution. There was a large amount of shrinkage in the outcome because the predicted value was pulled to the left of the sample

value. This is because the Study 2 sample value was a bit extreme compared to the other study values.

In summation, METAR detections during flight performance were the same for salience enhanced D1 and D3 symbols. There was no difference in detection performance between triangles (D1) and circles (D3). A meta-analysis showed that the overall effect of an increase in METAR symbol salience for VFR-to-IFR changes improved detection performance by an estimated 12%. This is a small effect, and we therefore believe that METAR symbol color changes are not efficient enough to attract pilot attention. Other more efficient change notifications need to be coupled to METAR color changes, such as the vibrotactile bracelet investigated by Ahlstrom et al. (2015b).

2.2.4 The Use of AWOS Information During Flight

During simulation flights, pilots could tune in to selected AWOS stations and receive weather updates. There were seven AWOS stations available during the scenario: Martinsburg (KMRB); Baltimore (KBWI); Washington, DC (KDCA); Easton/Newman (KESN); Washington Dulles (KIAD); Johnstown (KJST); and Hagerstown (KHGS).

At the start of the scenario all AWOS stations reported VFR conditions with visibilities ranging from five to 10 miles, and sky conditions with scattered clouds between 4000–9000 feet and overcasts between 8000–10,000 feet.

However, there was a reduction in the scenario visibility over time. After the first METAR change (VFR-to-IFR) at KMRB (10 min), the AWOS reported a 1-mile visibility at Martinsburg airport. Similarly, at the 19-minute VFR-to-IFR METAR changes at the KBWI, KDCA, KESN, KIAD, and KJST airports, these AWOS stations reported visibilities in the range of 1–2 miles. Finally, at the 30-minute VFR-to-IFR METAR change at KHGS, the AWOS reported 1-mile visibility.

Ideally, pilots would use the METAR color changes with the AWOS information to maintain their WSA. This would enhance their decision-making by providing important information that could be used for decisions to deviate to an alternate airport, turn around, or request an IFR flight plan (for IFR-rated pilots).

During the scenario, we recorded each time pilots tuned in to an AWOS station. We then summarized these count values for each pilot and group (D1 and D3). Figure 13 shows the total number of AWOS counts. The number of AWOS inquiries was 143 and 165 for D1 and D3, respectively. However, not all pilots used the AWOS information during flight. For D1, only 33 of the 46 pilots (72%) tuned in to one or more AWOS stations. For D3, 39 of the 47 pilots (83%) used AWOS information.
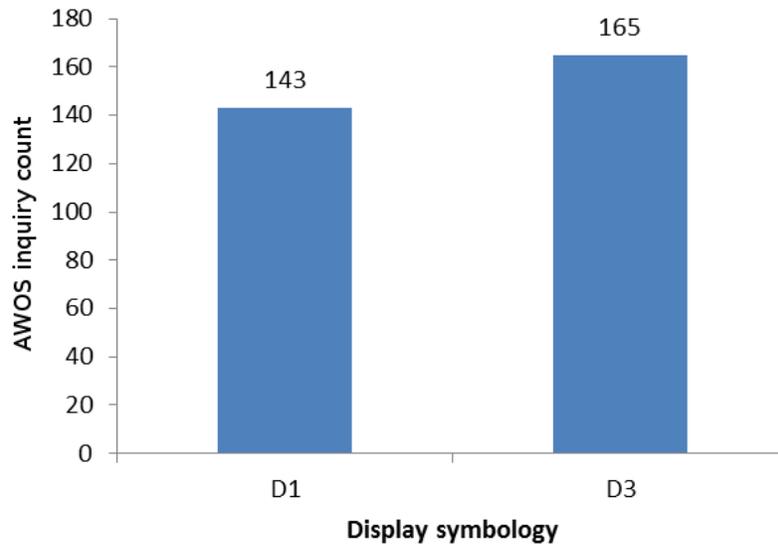
Figure 13. The total number of AWOS inquiries for D1 and D3.

Although the AWOS inquiry count is higher for the D3 group than the D1 group, this difference is marginal. Using a Bayesian beta-binomial model with MCMC sampling, we find that the D1 versus D3 difference is not credible (see Figure 14) as the mean posterior difference is -0.07 with the value 0 included in the 95% HDI.
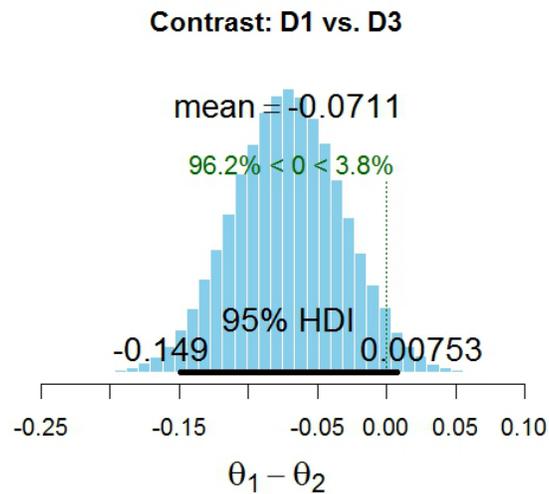
Figure 14. Posterior contrasts for the difference in AWOS counts between D1 and D3.

We also assessed the AWOS usage across the seven AWOS stations. Figure 15 shows the AWOS stations (red dots) and their relation to KABE (Allentown, scenario start point) and KMRB (Martinsburg, destination airport).
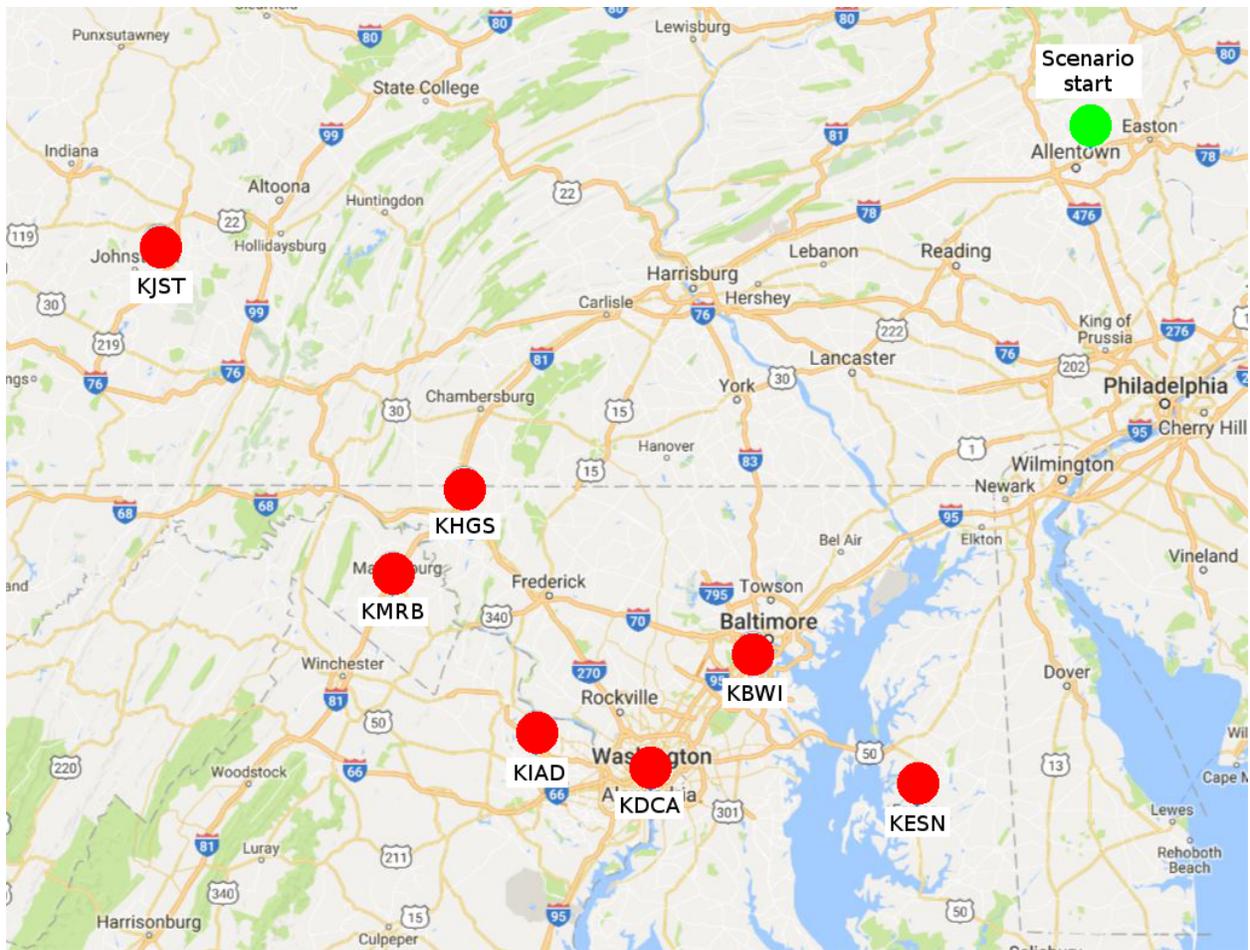
Figure 15. The seven AWOS stations (red dots) and their relation to the scenario start point at KABE (Allentown, green dot) and the destination at KMRB (Martinsburg).

Figure 16 shows the scenario time for each AWOS inquiry. Of the seven AWOS stations, none of the pilots tuned in to the AWOS station at KJST (Johnstown). For the remaining six stations, the AWOS counts are highest for BWI (74), HGS (93), and MRB (117). Furthermore, pilots tuned in to these stations during the entire scenario. For the remaining three stations—DCA (5), ESN (11), and IAD (8)—AWOS inquiries were much less frequent and roughly occurred after 20 minutes into the scenario. Using a model that accounts for groups (D1 and D3) and AWOS stations, we found no credible differences in counts between D1 and D3 for any of the six AWOS stations.
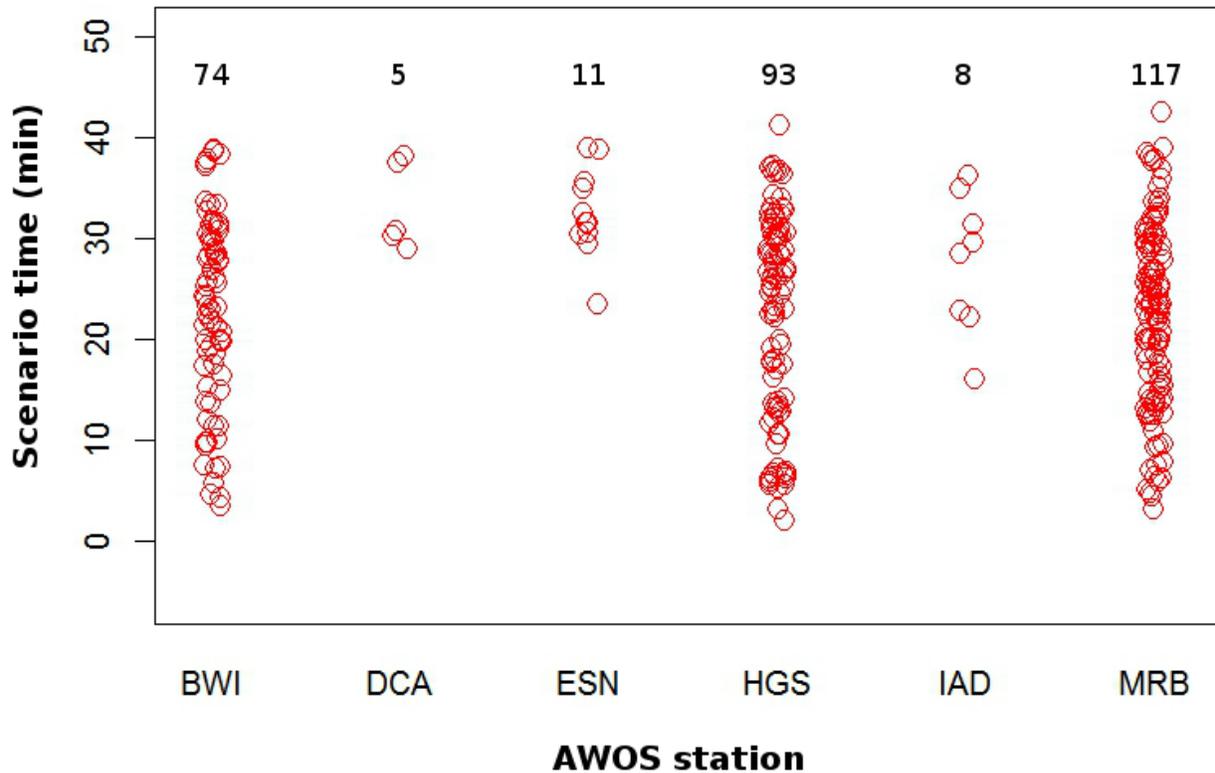
Figure 16. The scenario time for AWOS inquiries (red circles) at six AWOS stations. The numbers above the data columns represent the total count of AWOS inquiries.

In summation, AWOS stations provide pilots with important information regarding visibility and sky conditions at airports. This information could increase pilot WSA and enhance pilot decision-making whether to turn around, deviate, or continue to the destination airport. However, we found that for D1 and D3 only 72% and 83% tuned in to one or more AWOS stations, respectively. The destination airport at MRB received the highest count of AWOS inquiries, followed by the nearby station at HGS, and finally the AWOS station at BWI.

2.2.5 Out-the-Window Visibility Reports

During the SAGAT stops at the 11-, 20-, and 35-minute marks, we requested pilots to provide an estimate of the out-the-window visibility. In this section, we provide the result of these visibility estimates. However, before analyzing the visibility data we performed an altitude analysis at the three reporting times to determine if there were any differences based on altitude. For this analysis we used the aircraft altitude for each visibility estimate and the factors display symbology (D1 and D3) and reporting order (i.e., the 11-, 20-, and 35-minute marks labeled as First, Second, and Third, respectively).

To analyze the data, we used a model from Kruschke (2014) for a metric predicted variable (i.e., altitude) with multiple nominal predictors (i.e., display symbology and reporting order). In the model, the group data were modeled as a random variation around an overall central tendency (baseline). The group data characteristics, like the group central tendency, were analyzed as a deflection from the baseline with the requirement that deflections sum to zero.

Figure 17 shows the aircraft altitude (red circles) for D1 and D3 at the First, Second, and Third visibility reporting probes. As is shown by the altitude data in Figure 17, the central tendencies only show a small reduction in altitude across the reporting times, with the highest altitudes for First, followed by Second, and the lowest altitudes for Third.
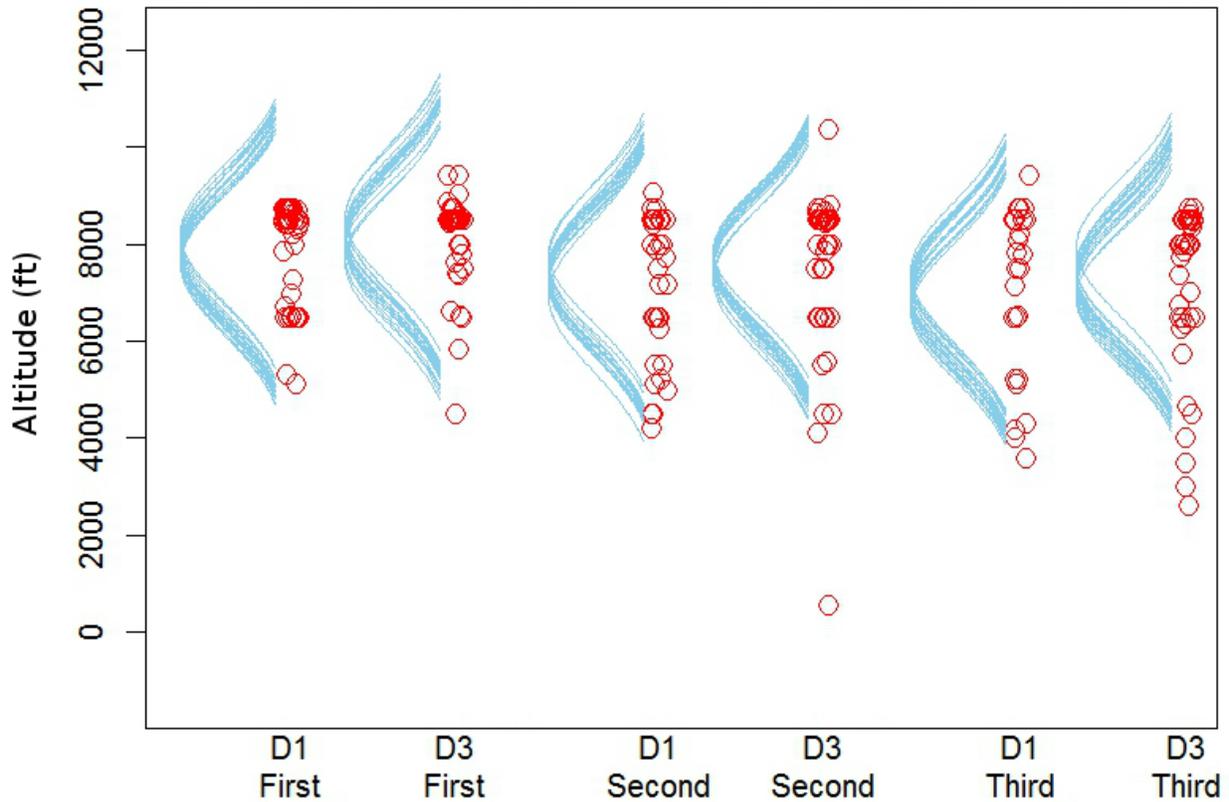
Figure 17. Individual aircraft altitudes (red circles) for D1 and D3 at the First (11 min), Second (20 min), and Third (35 min) visibility estimation times. The blue lines represent a posterior predictive check of how well the model fits the data.

Figure 18 shows the posterior contrasts for the difference in altitude between D1 and D3 at the three reporting times. The comparison between the altitudes for D1 and D3 at First (left) has a mode of -252 feet, with D3 having a higher altitude than D1. However, because the value 0 is included in the 95% HDI, this difference is not credible. The comparison between D1 and D3 at Second (middle) has a mode of -239 feet, implying that D3 pilots were at higher altitudes than D1 pilots were. However, the 95% HDI for the posterior difference distribution includes the value 0, so this difference is not credible. Finally, the comparison of altitudes for D1 and D3 at Third has a mode of -212 feet, implying that the D3 pilots were at higher altitudes than the D1 pilots were. Again, this difference is not credible because the 95% HDI includes the value 0.

D1 - D3 @ First

mode = -252

88.6% < 0 < 11.4%

95% HDI
-779   186

D1 - D3 @ Second

mode = -239

83.7% < 0 < 16.3%

95% HDI
-742   255

D1 - D3 @ Third

mode = -212

78.3% < 0 < 21.7%

95% HDI
-762   353

Difference in Altitude
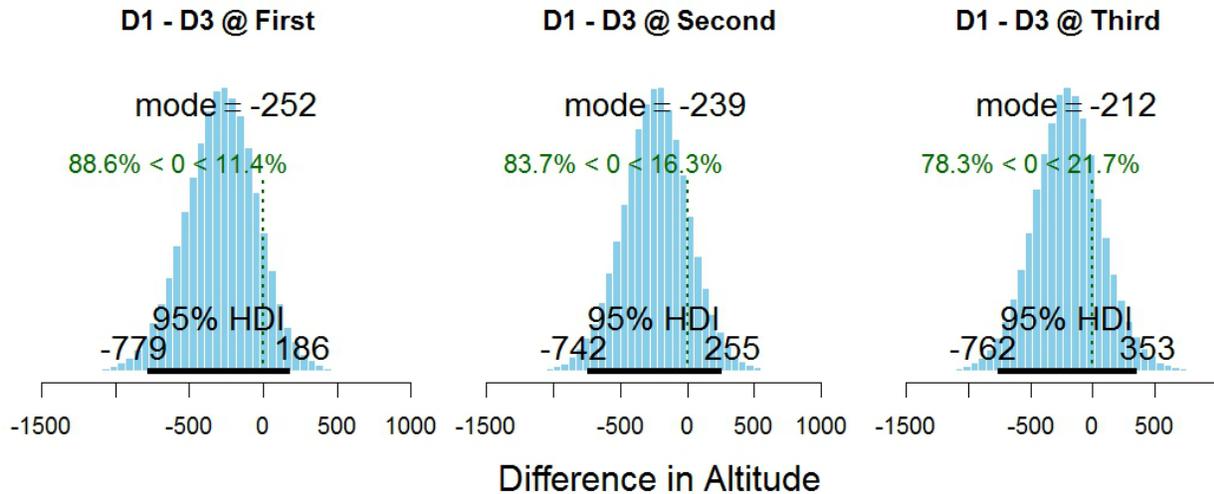
Figure 18. Posterior contrasts for the difference in aircraft altitudes between D1 and D3 at First (left), D1 and D3 at Second (middle), and D1 and D3 at Third.

Figure 19 shows pilot visibility estimates for D1 and D3 at the simulated 10-, 20-, and 30 nmi visibility zones. During continuous flight (excluding turning around or landing at alternate airports), pilots navigated within these simulated visibility zones during the SAGAT stops at the 11-, 20-, and 35-minute marks.

The red circles in Figure 19 represent individual visibility estimates for the six combinations of display symbology and visibility zone. The horizontal green lines represent a perfect correspondence between the reported and simulated visibilities. As we can see in the figure, there is an underestimation of the reported visibility for all simulated visibility zones.
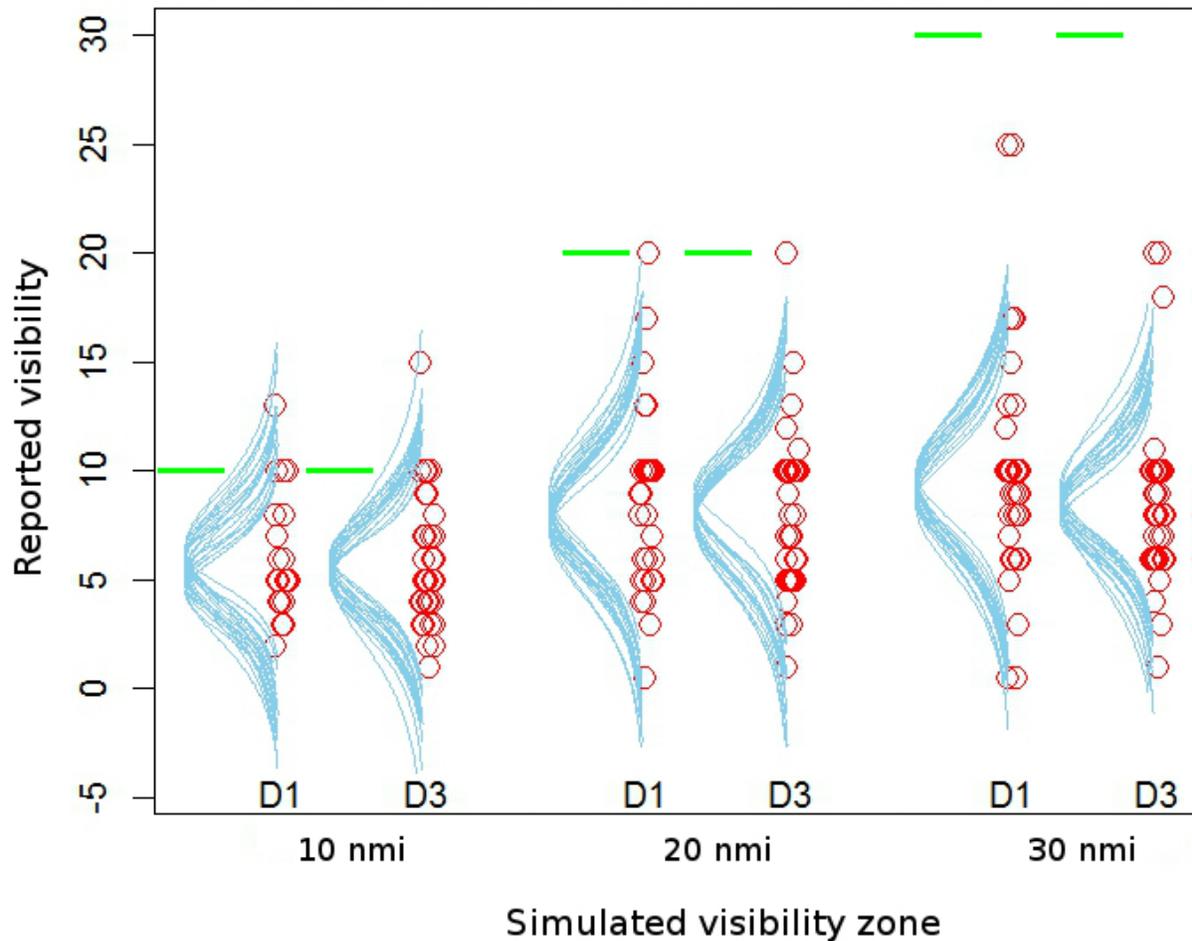
Figure 19. Individual visibility estimates (red circles) for D1 and D3 at the 10-, 20-, and 30 nmi visibility zones. The blue lines represent a posterior predictive check of how well the model fits the data.

We were interested in assessing the mean reported visibilities for D1 and D3 at the three visibility zones, and whether there was a difference in visibility reports between D1 and D3.

As we can see in Figure 19 (left), the reported visibilities for the D1 and D3 at the 10 nmi visibility zone are very similar. The majority of the reported visibility estimates are well below the simulated visibility, denoted by the green line. Only a few reported visibilities are at the simulated visibility. The model analysis predicted a mean posterior visibility for D1 and D3 of 5.5 (95% HDI from 4.5 to 6.5) and 5.7 (95% HDI from 4.8 to 6.6) miles, respectively.

For the 20 nmi visibility zone there is an even greater underestimation of the simulated visibility. As is shown in Figure 19 (middle), only a few reported visibility estimates coincide with the simulated visibility. The majority of the reported visibilities are at 10 miles or below. The posterior mean for D1 was 8.2 miles with a 95% HDI from 7.3 to 9.2. For D3, the model predicted a mean of 8 miles with a 95% HDI from seven to 8.9.

The visibility estimate data for the 30 nmi visibility zone (Figure 19, right) is similar to the data for the 20 nmi visibility zone. However, there is an even greater underestimation in the reported

visibilities for the 30 nmi zone compared to the 20 nmi zone. None of the reported visibilities for the 30 nmi zone were above 25 miles, and the majority of the individual estimates were at or below 10 miles. The output from the analysis shows a mean posterior visibility of 9.1 miles for D1, with a 95% HDI from 8.2 to 10. For D3, the mean predicted visibility is 8.6 miles, with a 95% HDI from 7.8 to 9.3 miles.

Figure 20 shows the posterior contrasts for a comparison of the visibility reports at the 10-, 20-, and 30 nmi visibility zones and the comparison of reported visibilities between D1 and D3. The contrasts reflect the underestimation and data trends reported above. Pilots reported credibly greater visibilities for the 20 nmi visibility zone compared to the 10 nmi zone (mode of difference = -2.53), and greater visibilities for the 30 nmi zone compared to the 10 nmi zone (mode of difference = -3.27). Pilots did not report credibly higher visibilities for the 30 nmi zone compared to the 20 nmi zone (mode of difference = -0.67 and the value 0 included within the 95% HDI). Furthermore, there was no credible difference in visibility reports between D1 and D3 (mode of difference = 0.18 and the value 0 included within the 95% HDI).
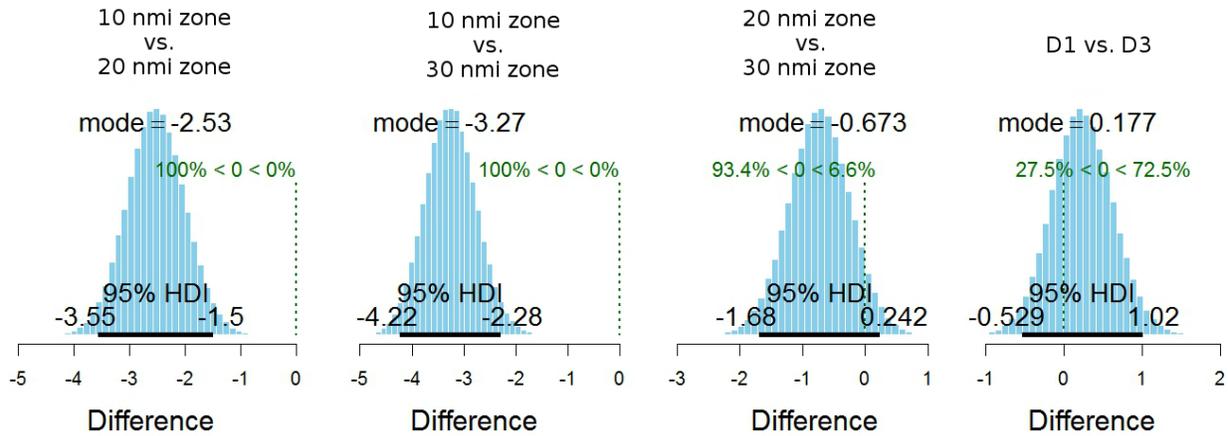
Figure 20. Posterior contrasts for comparisons of reported visibility between visibility zones and display symbology (D1 and D3).

We assessed the altitudes and reported visibilities for D1 and D3 at three time marks during the simulation flight. Although we found no credible differences in pilot altitudes, there were large variations among individual pilots in both altitude and visibility estimates. This brings up the question of the relationship between aircraft altitude and reported visibility. Potentially, the large variation in the reported visibilities could be due to the variation in aircraft altitudes (i.e., differences in the slant-range visibility).

To assess this possibility, we used the altitude (x) and corresponding visibility estimates (y) in a robust linear regression model by Kruschke (2014). In the model, each predicted y value is computed as $y = \beta_0 + \beta_1 x$ where β_0 is the y-intercept (where the regression lines intersect the y-axis when $x = 0$) and β_1 is the slope (indicates how much y increases when we increase x by 1). To be robust against outliers, the model uses a t-distribution for the noise distribution instead of a normal distribution (i.e., Gaussian distribution). At the lowest level of the model, each datum comes from a t-distribution with a mean μ, a scale parameter (i.e., standard deviation) σ, and a normality parameter v. The prior on the scale parameter is a broad uniform distribution, and the normality parameter v has a broad exponential prior. Both β_0 and β_1 have broad normal priors that are noncommittal and vague on the scale of the data.

If the aircraft altitude determined the reported visibility in our scenario, we expected to see a linear relationship between altitude (x) and reported visibility (y). Conversely, if there was no relationship between the aircraft altitude and reported visibility, we expected to see an estimated posterior slope of zero.

Figure 21 shows the outcome of the regression analysis with aircraft altitude on the x-axis and reported visibility on the y-axis. The black circles are the data points—one altitude value per pilot visibility estimate. The lines represent regression lines and the three vertical distributions are superimposed *t*-distributed noise distributions. As shown by the regression lines in Figure 21, aircraft altitude does not seem to have a large effect on the reported visibilities. On the contrary, the reported visibilities for aircraft at the same altitude have a wide distribution.
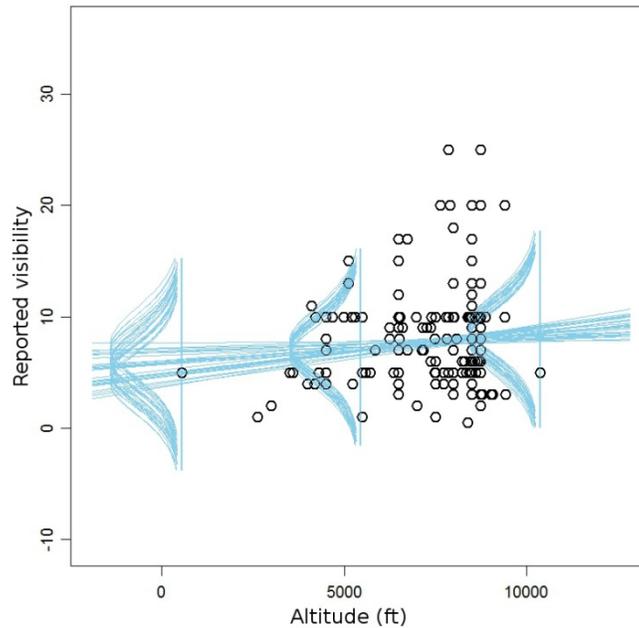
Figure 21. Regression lines and noise distributions for the prediction of reported visibility from aircraft altitude.

Figure 22 shows the mean posterior outcome for β_0 (intercept), β_1 (slope), and the scale parameter σ. The intercept has a mode of 5.93, which is the predicted value of *y* (visibility) when *x* (altitude) = zero. The non-credible slope (the value 0 is located within the 95% HDI) has a posterior mode of 0.0002. This means that as we increase altitude by one, visibility will increase by the value of β_1 (i.e., 0.0002). This outcome implies that aircraft altitude did not determine the reported simulation visibilities.
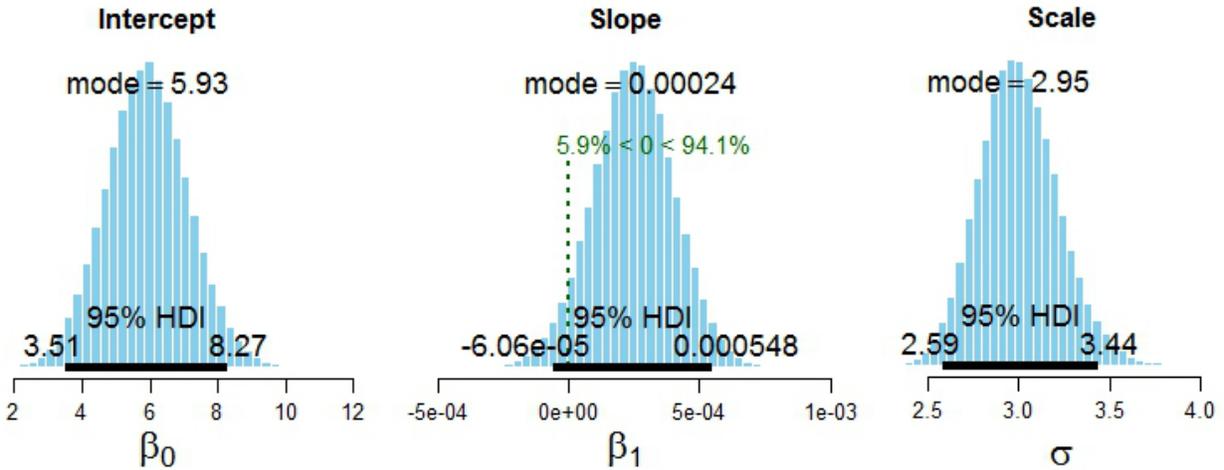
Figure 22. The posterior intercept (β_0), slope (β_1), and scale (σ) parameters from the robust linear regression analysis.

In summation, pilots provided visibility reports at three times during the simulation. We correlated these reported visibilities with the simulated visibilities in three visibility zones (i.e., 10, 20, and 30 nmi). We found that pilots greatly underestimated the simulated visibility in all visibility zones. By regressing the reported visibilities on aircraft altitude, we determined that the reported visibilities were not a function of aircraft altitude. These results reveal a gap in ability of pilots to determine out-the-window visibility accurately. Although the estimation errors imply that visibility judgments were conservative, they reveal an inaccurate assessment on the part of the pilots, which speaks against a high WSA during flights in deteriorating visibility.

2.2.6 Decision-Making

In this section, we present results from pilot decision-making during flight. Specifically, we analyzed the number of pilots who continued flight toward the destination airport, diverted to an alternate airport, or requested an IFR flight plan. During our analysis, we also correlated the timing of each pilot decision with the aircrafts' geographical location. This allowed us to determine the simulated visibility zone at each decision point.

Figure 23 shows the scenario visibility zones and summarizes the decision data by area of decision point. The table insert provides a summary of the areas of simulated visibility and the total number of pilots that continued into each of the respective areas. We used the flight track data to determine the area farthest traveled by each pilot. For each area of visibility, the table also shows the median age of the pilots; the number of IFR-rated pilots; number of non-rated pilots; median total flight hours; number of pilots in the D1 and D3 conditions; and the number of pilot decisions to divert, continue, or request an IFR flight plan.

Vis Range	Total Pilots	Age	IFR rated	no rating	Med Flight Hours	D1	D3	Divert	Cont'd	IFR PopUp
2nm	-	-	-	-	-	-	-	-	-	-
4nm	-	-	-	-	-	-	-	-	-	-
8nm	20	49	11	9	313	9	11	1	15	4
10nm	53	48	20	30	450	23	30	31	6	16
20nm	19	52	7	11	297	13	6	17	1	1
30nm	1	27	0	1	100	1	0	1	0	0
50nm	-	-	-	-	-	-	-	-	-	-
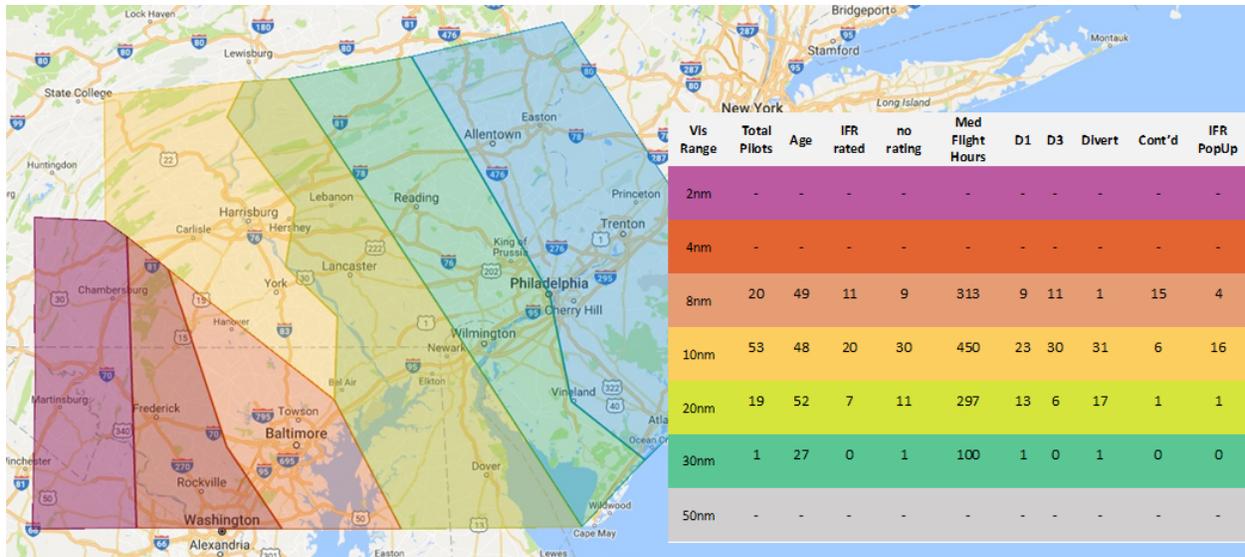
Figure 23. Pilot decisions to continue flight, divert, or file an IFR flight plan ("IFR PopUp") at four simulated visibility zones.

We used the same model for count-valued data (i.e., contingency table analysis) to analyze these data as we used in the previous AWOS result section.

As shown in Figure 23, there was only one pilot who decided to divert in the 30 nmi visibility zone. For the 20 nmi visibility zone, there were 19 pilot decisions (see Figure 24). Seventeen (90%) of these were pilot decisions to divert. There was only one decision to continue and to request an IFR flight plan, respectively. There was no credible difference between D1 and D3 (mode of difference = 1.06, 95% HDI from -0.62 to 5.6).
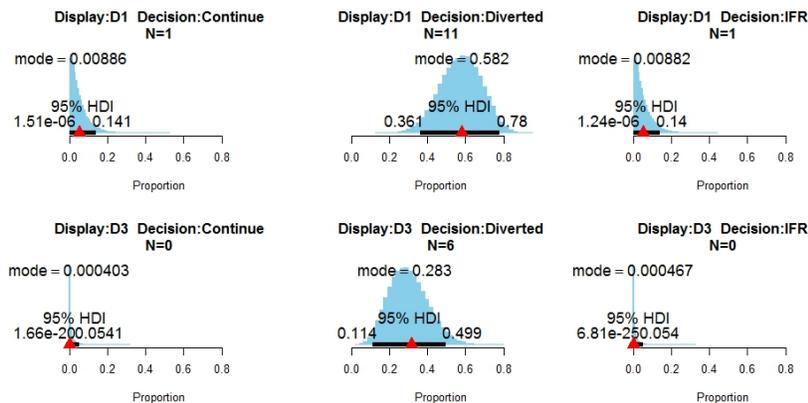
Figure 24. The actual counts and posterior distributions of estimated cell proportions for pilot decisions to continue, divert, and request IFR flight plans in the 20 nmi visibility zone. The red triangles indicate the actual data proportions.

For the 10 nmi visibility zone, there were 53 pilot decisions (see Figure 25). Six of these were decisions to continue flight (11%); 31 to divert (59%); and 16 to request an IFR flight plan (30%). Overall, there were credibly more pilot decisions to divert than decisions to continue

flight (mode of difference = 1.56, 95% HDI from 0.78 to 2.6). There were also credibly more decisions to divert than to request an IFR flight plan (mode of difference = 0.71, 95% HDI from 0.1 to 1.4). However, there was no credible difference between continue flight versus requesting an IFR flight plan (mode of difference = -0.9, 95% HDI from -1.91 to 0.007) and no credible difference between D1 and D3 (mode of difference = -0.27, 95% HDI from -0.95 to 0.33).
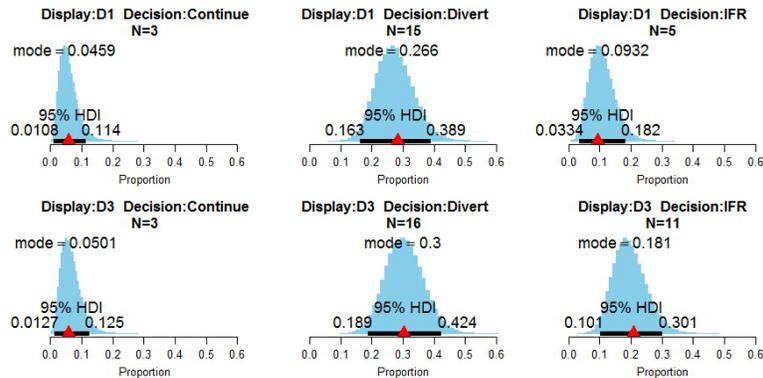
Figure 25. The actual counts and posterior distributions of estimated cell proportions for pilot decisions to continue, divert, and request IFR flight plans in the 10 nmi visibility zone. The red triangles indicate the actual data proportions.

Finally, for the eight nmi visibility zone (see Figure 26), most decisions were to continue flight toward the destination airport (75%) with one decision to divert and four decisions to request an IFR flight plan. Whereas the frequency of decisions to continue versus decisions to divert (mode of difference = -2.48, 955 HDI from -5.92 to -0.85) and the frequency of decisions to continue versus decisions to request and IFR flight plan (mode of difference = 1.39, 95% HDI from 0.32 to 2.84) were credibly different, there was no credible difference between D1 and D3 (mode of difference = -0.109, 95% HDI from -1.5 to 1.8).
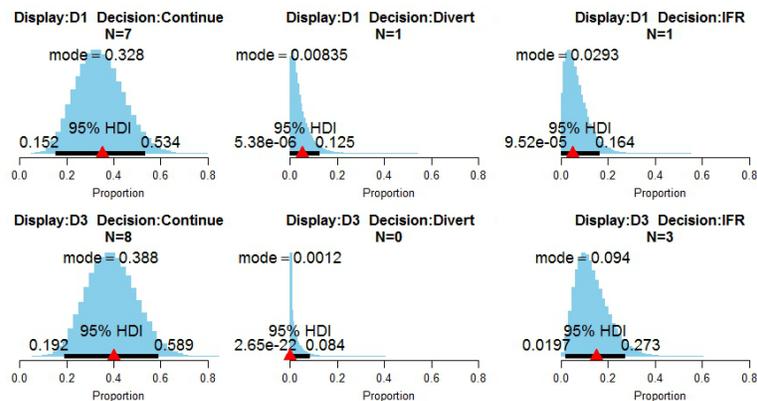
Figure 26. The counts and posterior distributions of estimated cell proportions for pilot decisions to continue, divert, and request IFR flight plans in the eight nmi visibility zone. The red triangles indicate the actual data proportions.

In summation, we analyzed pilot decisions to continue flight, divert, or request an IFR flight plan during flight. These decision points correspond to flights into four different visibility zones.

The outcome shows that only one pilot made a flight decision in the 30 nmi visibility zone to divert to an alternate airport. The majority of pilots who made a flight decision (90%) in the 20 nmi visibility zone opted to divert to an alternate airport. For pilot decisions in the 10 nmi visibility zone, the majority of decisions were to divert to an alternate airport (59%) and to request an IFR flight plan (30%). For the eight nmi visibility zone, the majority of decisions were to continue flight toward the destination airport (75%). Finally, there were no credible differences in pilot decision points between D1 and D3.

2.3 Summary of Findings for Simulation 1

1. An analysis of pilot/ATC communication contents revealed a similar communication count for the D1 and D3 groups. We found the highest counts for AWOS communications and the lowest counts for IMC conditions and IFR requests, with altitude and turn around-divert communications in between.
2. For METAR detections during flight, performance was the same for salience enhanced D1 and D3 symbols. There was no difference in detection performance between triangles (D1) and circles (D3).
3. A meta-analysis showed that the overall effect of an increase in METAR symbol salience for VFR-to-IFR changes only improves detection performance by an estimated 12%, at best. Therefore, we believe that METAR symbol color changes need to be coupled with other sensory modalities (e.g., vibrotactile notifications).
4. AWOS stations provide pilots with important information regarding visibility and sky conditions at airports. However, we found that for D1 and D3 only 72% and 83% of the pilots tuned in to one or more AWOS stations, respectively.
5. We found that pilots greatly underestimated the simulated visibility in all visibility zones. This result reveals a gap in pilots' ability to determine out-the-window visibility accurately and implies a low WSA during flights in deteriorating visibility.
6. The majority of pilots who made a flight decision (90%) in the 20 nmi visibility zone opted to divert to an alternate airport. For pilot decisions in the 10 nmi visibility zone, the majority of decisions were to divert to an alternate airport (59%) and to request an IFR flight plan (30%). For the eight nmi visibility zone, the majority of decisions were to continue flight toward the destination airport (75%).

3. SIMULATION 2

The purpose of Simulation 2 was to evaluate GA pilot behavior and decision-making under VFR flights in deteriorating weather. We used a scenario based in Alaska in which half of the pilots (experimental group) were equipped with a portable weather application that provided an AR for precipitation and visibility (forecasts) information.

We hypothesized that pilots in the experimental group would exhibit improved decision-making and WSA compared to pilots flying without the AR application (control group).

3.1 Method

3.1.1 Participants

Ninety-five GA pilots participated in Simulation 2. These pilots came from the same participant pool as the pilots for Simulation 1. Like Simulation 1, each pilot was randomly allocated to one of two simulation groups (i.e., experimental or control). We present the pilot background information in Table 7.

Table 7. Descriptive Characteristics of Study Participants

Group	n	Age (years)		Flight hours accrued					
				Total		Instrument		Instrument- last 6 mo.	
		Median	Range	Median	Range	Median	Range	Median	Range
Control	52	55	18-83	450	78-4500	15	0-350	0	0-89.6
Experimental	47	48	19-82	300	65-5000	10	0-1000	4	0-70

3.1.2 Aircraft Simulator

For Simulation 2, we used a single-engine Redbird C172 (motion base) with a 180° out-the-window view (see Figure 27). For flight dynamics, the Redbird simulator used Microsoft® Flight Simulator X®. To meet our weather display needs, we used X-Plane 10 to drive the six out-the-window view monitors. The simulator was equipped with a stand-alone portable weather display running on a Windows Surface Pro 3 and a voice communication system that provided a link between the pilots and ATC through a PTT capability.

Figure 27. Left: an exterior view of the Redbird C172 fuselage. Right: the cockpit out-window view, the G1000-type GA glass cockpit control display, the instruments, and the stand-alone weather display running on a Windows Surface Pro 3.

3.1.3 Flight Plan

The VFR flight departed from a location northwest of Juneau (starting at altitude, 6000 ft.), Alaska (PAJN), with the destination at Skagway airport (PAGY) as shown in Figure 28. The flight progressed through a narrowing pass and continued into a narrowing fjord with gradually reduced visibility. The Alaska scenario highlighted the hazards of flying in deteriorating weather when the terrain (canyons) and low ceilings present few alternatives for turning around or flying toward an alternative airport. The destination Skagway has a near sea level runway situated in a

narrow valley. There are also steep mountains on either side of the route, which forces the airplane to fly along a narrowing fjord.
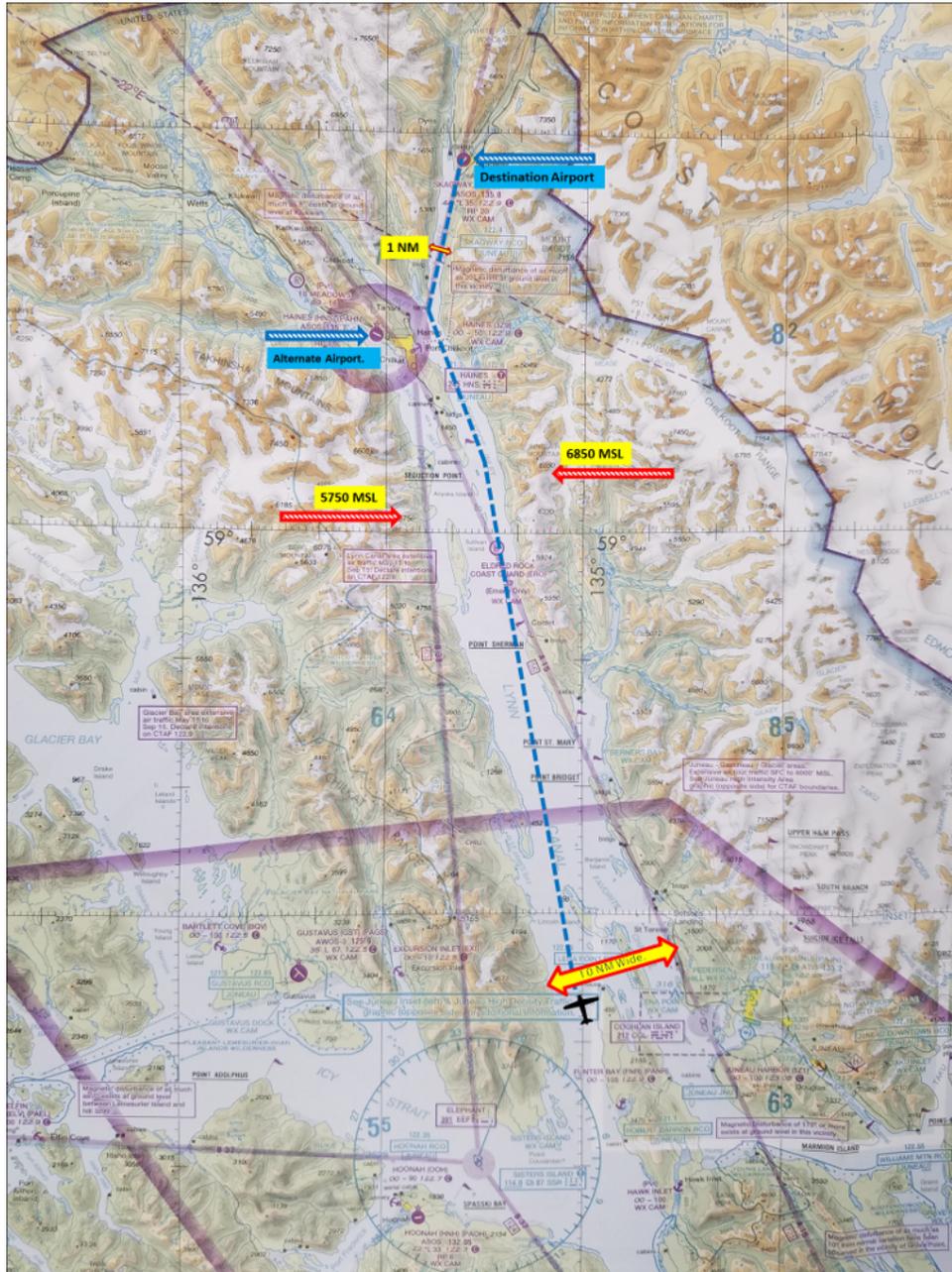
Figure 28. The Alaska scenario route used during Simulation 2.

3.1.4 Weather Scenario – Reductions in Visibility

To assess pilot WSA during flight, we created an out-the-window scenario in which the visibility decreased as pilots flew toward the destination airport. Similar to Simulation 1, we used METAR strings to create six different visibility areas (see Table 8).

Table 8. Visibility METAR Strings

Visibility	Identifier	METAR String
25nmi	PAGS	102100Z 26306KT 8SM +TSRA OVC000TCU A2992
12 nmi	MDEG	135.2293 58.9769 102100Z 26306KT 12SM OVC120 A2992
8nmi	MDEG	135.2856 59.1370 102100Z 26306KT 8SM OVC100 A2992
6nmi	MDEG	135.3845 59.2425 102100Z 26306KT 6SM OVC100 A2992
4nmi	MDEG	135.3790 59.3196 102100Z 26306KT 4SM OVC100 A2992
1 nmi	MDEG	135.3516 59.4120 102100Z 26306KT 1SM OVC100 A2992

Visibility was 25 nmi at the start of the scenario, but the conditions along the route deteriorated and visibility was progressively reduced to 12 nmi, 8 nmi, 6 nmi, 4 nmi, and finally 1 nmi at the destination airport (shown in Figure 29).
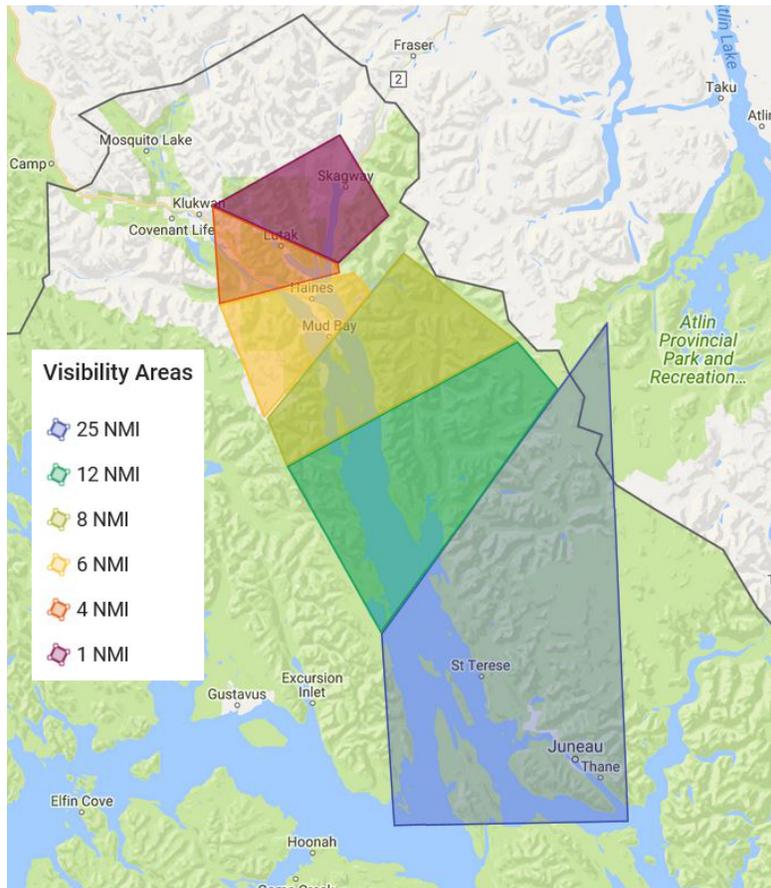
Figure 29. Simulated visibilities along the route from start (west of Juneau) to destination (Skagway airport).

Besides the simulated visibility reductions, the scenario contained a line storm to the west of the scenario start position. However, this storm was stationary and never moved into close proximity of the scenario route.

3.1.5 Weather Presentation

In this simulation, half of the pilots were equipped with a portable weather application during flight. The application was mounted inside the cockpit and displayed a Google Earth map background, the route (purple line), range rings, aircraft position symbol, and areas of precipitation (see Figure 30). Besides the precipitation information, the application could also display an overlay of forecasted visibility conditions. These data are similar to the graphical visibility charts found on the Aviation Weather Center's Aviation Digital Data Service web page: https://www.aviationweather.gov/adds/.

To support the pilot, the application used an AR that automatically displayed a blue line from the aircraft position symbol to an area of potentially hazardous weather. We chose the blue line because of its high salience and the fact that it yielded the best detection performance in a change-detection experiment (Ahlstrom & Racine, 2019). For this simulation, the blue AR line appeared as soon as the aircraft was 20 nmi from a precipitation area or 20 nmi from an area where the forecasted visibility was three nmi. The AR logic has previously been explored for the generation of automated alerts and weather advisories in the ATC environment (Ahlstrom, 2015; Ahlstrom & Jaggard, 2010). Essentially, the AR logic keeps track of hazardous weather locations and reminds the pilot when a specific threshold is reached. For simulation purposes, we used 20 nmi for precipitation and visibility thresholds during all simulation runs. Real-world AR applications, however, would allow the pilot to set the distance parameters prior to flight, or to change the parameters in flight as needed.
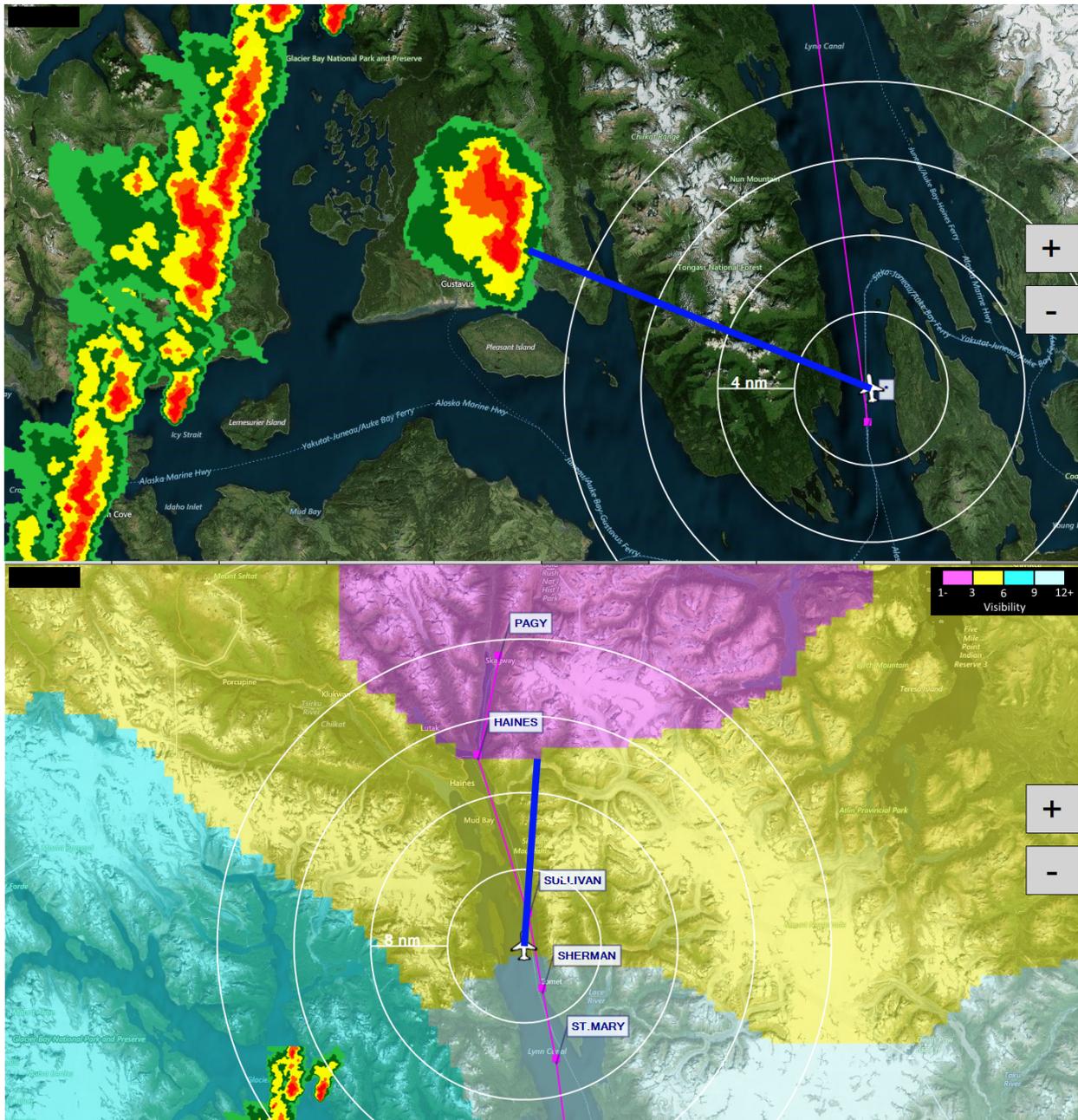
Figure 30. The weather application display of precipitation areas (left), areas of forecasted visibility (right), display zoom ("+" and "-"), and the blue line AR.

3.1.6 Simulation Procedure

This was the same as Simulation 1 except for the content of the pre-flight briefing, the added prompts for flight conditions (i.e., controller queried pilot for altitude and visibility estimates), the post-run questionnaire (see Appendix H), and the fact that no SAGAT method was used.

3.1.7 Independent Variable

The independent variable in Simulation 2 was weather presentation (control group with no weather application and no AR versus the experimental group with a weather application and an AR).

3.1.8 Dependent Variables

During Simulation 2, we recorded dependent variables that captured the following categories: system performance, communication, WSA, and decision-making. Of particular interest for Simulation 2 was the comparison between the experimental group and the control group regarding the effect on pilot behavior and decision-making from the AR display.

During the simulation we measured the distance-to-weather (≥ 30 dBZ precipitation cells) once every five seconds. In addition to the distance-to-weather measures, we probed pilots with detailed post-scenario questions regarding distances to precipitation cells and flying behavior (see Appendix H). This allowed us to acquire data for an assessment of pilot WSA and the perceived distance from the aircraft to hazardous cells and whether pilots were cognizant of their closest point of approach. It should be noted, however, that for this simulation we did not brief pilots about current recommendations to stay at least 20 statute miles away from storm cells. In Table 9, we provide a list of the dependent variables and a short description.

Table 9. Dependent Variable List

Number	Dependent variable	Description
1	System performance measures	Data from the cockpit simulator (e.g., altitude, heading, lat/long position).
2	Pilot/ATC communications	The content and duration of pilot/ATC communications.
3	WSA	Pilot response to ATC weather requests (i.e., pilot visibility estimates at 6 min, 12 min, and 19 min into the scenario).
		Pilot use of Automated Surface Observing System (ASOS) stations.
		Closest distance from aircraft to 30 dBZ precipitation areas.
4	Decision-making	Pilot decision to turn around, deviate, or to continue flight to the destination airport at Skagway.

3.2 Results

3.2.1 ATC Communication

Similar to Simulation 1, we also summarized researcher communication notes and categorized pilot communications into five categories. The first category, landing, includes communications that are related to pilot questions or statements about landing. The second category, ceiling-visibility-fog, includes communications about pilot reports of deteriorating weather conditions. The third category, altitude change, includes communications about pilot altitude requests. The fourth category, turn around-divert, includes pilot communications regarding decisions or questions about turning around or deviating to an alternate airport. Finally, the fifth category, AWOS, includes communications about going off the frequency to tune in to one or more AWOS stations.

Figure 31 shows a summary of the communication frequencies (i.e., count) for each of the five categories.
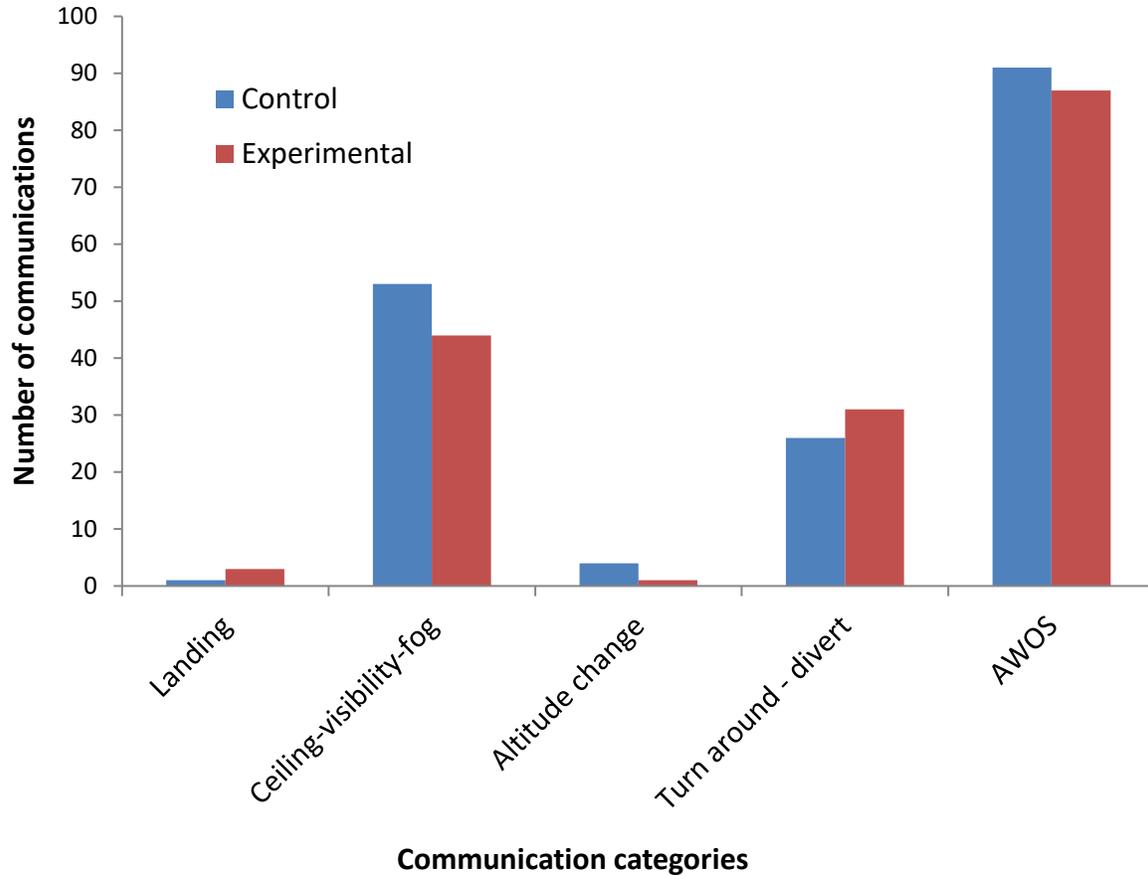
Figure 31. Summary of pilot communications for five communication categories.

As is shown in Figure 31, the category counts are very similar for the experimental and control groups, with the highest counts for AWOS communications, the lowest counts for landing, and the number of ceiling-visibility-fog, altitude change, and turn around-divert communications in between. Most importantly, there are no large and important deviations in frequency counts between the experimental and control groups.

In summation, an analysis of the pilot/ATC communication contents reveals a similar communication trend for the experimental and control groups. We find the highest counts for AWOS communications, followed by ceiling-visibility-fog, turn around-divert, altitude change, and the lowest communication count for landing.

3.2.2 Weather Situation Awareness

In the following section, we present results from an assessment of pilot WSA. We used three different dependent measures to capture WSA during Simulation 2: the distance from the aircraft to 30 dBZ precipitation areas, the use of AWOS information, and pilot out-the-window visibility

reports. To complement these dependent measures, we used pilot answers to detailed post-scenario questions (see Appendix H).

3.2.3 Distance from Aircraft to Precipitation Areas

FAA and NOAA (1983) recommend that pilots avoid hazardous storm cells by at least 20 statute miles. However, previous research has shown that GA pilots in simulations often fly much closer than the current recommendations with a mean distance-to-weather ($\geq$ 30 dBZ precipitation cells) ranging from of 7–14 nmi (Ahlstrom et al., 2015a, 2015b; Ahlstrom & Dworsky, 2012). Whereas it seems that pilot use of cockpit weather displays yields larger buffer distances to hazardous weather, it does not seem to encourage pilots to keep distances that are even close to the recommended 20-mile distance. Potentially, the discrepancy between pilot use of modern weather displays and the less-than-ideal flying behavior could be due to a weak WSA.

During flights, only the experimental group had access to the weather display with the AR for the 20 nmi precipitation distance. Therefore, we predicted that the experimental group would have an increased WSA for the exact location of the cells and a more rigorous awareness of how close they flew to these cells. The control group, however, had no access to a weather display with an AR and could only estimate the location of potential storm cells through the out-the-window view.

As the first step of the analysis, we assessed the closest distance-to-weather for each pilot in the experimental and control group. For this analysis we used one data value (closest distance) per pilot and a Bayesian model (Kruschke, 2014) for a metric-predicted variable (i.e., closest distance to weather) for two groups (experimental and control). In the model, the data are described by t distributions rather than normal distributions. Each group has different parameters for the means with broad normal distribution priors. Each group also has separate parameters for the standard deviations, with broad uniform distribution priors. However, both groups share a common normality parameter (v) that controls the height of the t distribution tails. The prior on v has an exponential distribution, which gives equal opportunity to small values of v and larger values of v. The common normality parameter means that both groups' data can inform the estimate of v. For the current analysis, we set v to a small value, which means that the t distributions have heavy tails and can therefore accommodate outliers in the data (i.e., robust estimation).

Figure 32 shows the data and the outcome of the analysis with predicted means, standard deviations (i.e., scale), and effect size. The top right histograms show the actual data with a superimposed posterior predictive check. The mean posterior distances are shown on the left, with a mode of 17.9 miles for the control group and 17.4 miles for the experimental group. Albeit small, this difference of means is credible with a mode of -0.43 miles. The scale (i.e., *SD*) has a posterior mode of 0.87 for the control group and 0.80 for the experimental group. The difference in scales with a mode of -0.06 is not credible, as the 95% HDI includes the value 0. The normality parameter (v) has a posterior mode of 1.38, indicating a nearly normal distribution (any value of v greater than ~ 1.47 on the $\log_{10}$ scale represents a nearly normal distribution). Finally, there is a credible effect size (i.e., a standardized change) of the differences with a posterior mode of -0.52.
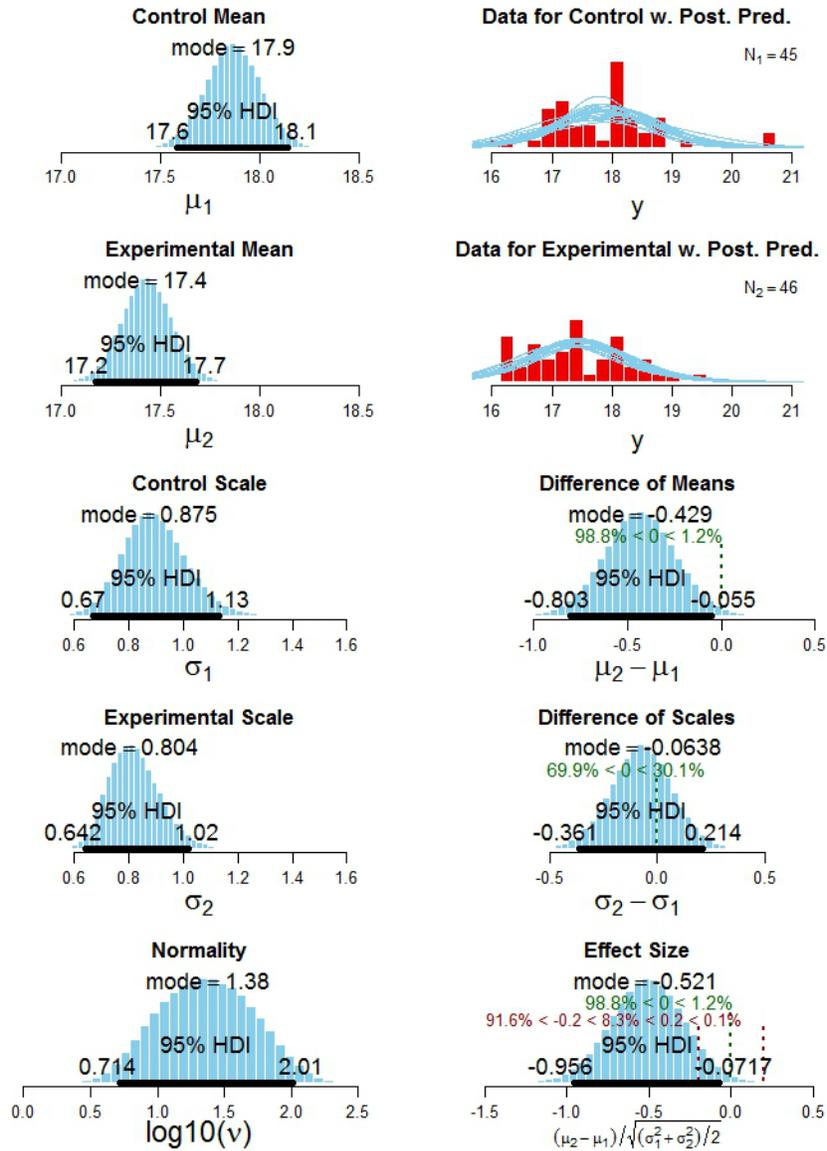
Figure 32. Comparison of the distance-to-weather data for the control and experimental groups (top right) with posterior distributions for means (top left), standard deviations (i.e., scale; middle left), difference of means and scales (middle right), effect size (bottom right), and the normality parameter (bottom left).

The initial distance-to-weather analysis revealed that pilots flew between 17.2 to 18.1 miles (95% HDIs) from precipitation cells. Next, we wanted to assess how cognizant pilots were of their closest point of approach.

To make this assessment we analyzed pilot answers to the post-scenario question: "At the beginning of the scenario, did you fly closer than 20 nautical miles (nmi) from the precipitation cells?" We correlated each pilots' "Yes" or "No" answer with the pilots' recorded closest point of approach. Figure 33 shows a summary of the number (i.e., counts) of yes and no answers (left) and the percentage correct answers (right) for the experimental group and the control group. Not included in these numbers are three pilots from the control group who stated that

they did not know whether they flew closer than 20 nmi or not, and therefore did not answer the question.

As expected, the experimental group had more correct (63%) answers than the control group (33%). Interestingly, all no answers for the experimental group were wrong, meaning all pilots did fly closer than 20 nmi to the storm cells. For the control group, there were two pilots above 20 nmi for their closest point of approach.
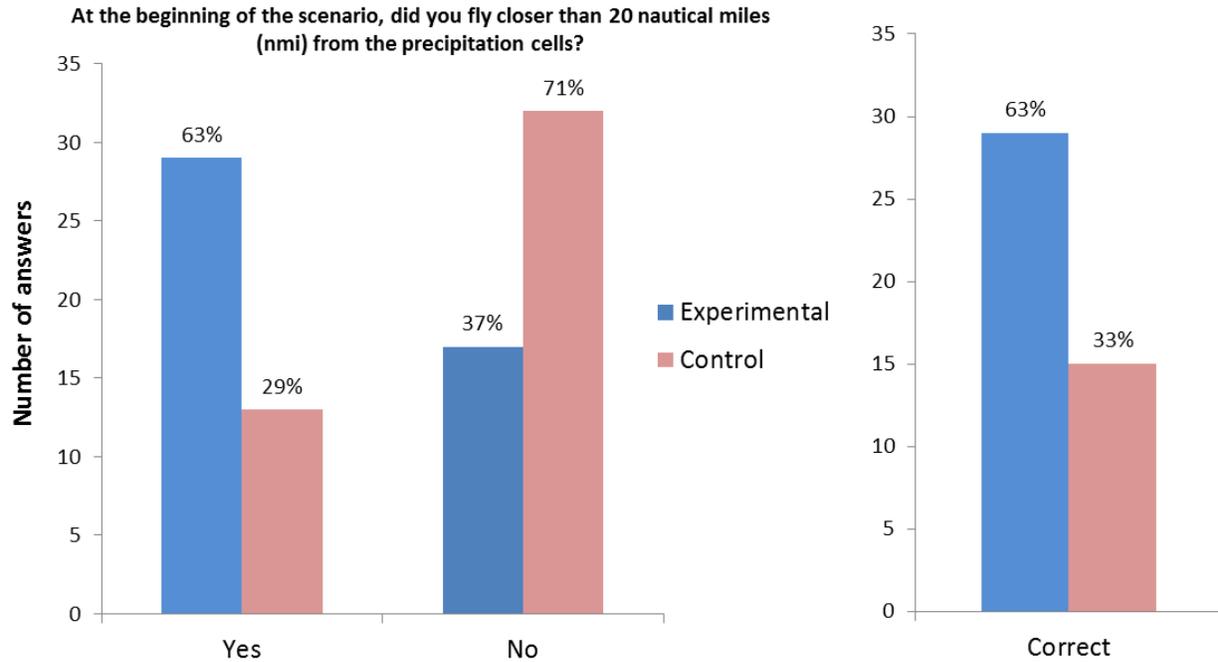
Figure 33. The number and percentage correct answers to the distance question.

To analyze the difference in the correct number of answers we used the same model from Kruschke (2011) for between-subjects analysis of dichotomous data that we used in Simulation 1. Figure 34 shows the posterior distributions as a function of the accuracy (μ_c) and precision (k_c) for correct answers to the post-scenario question.
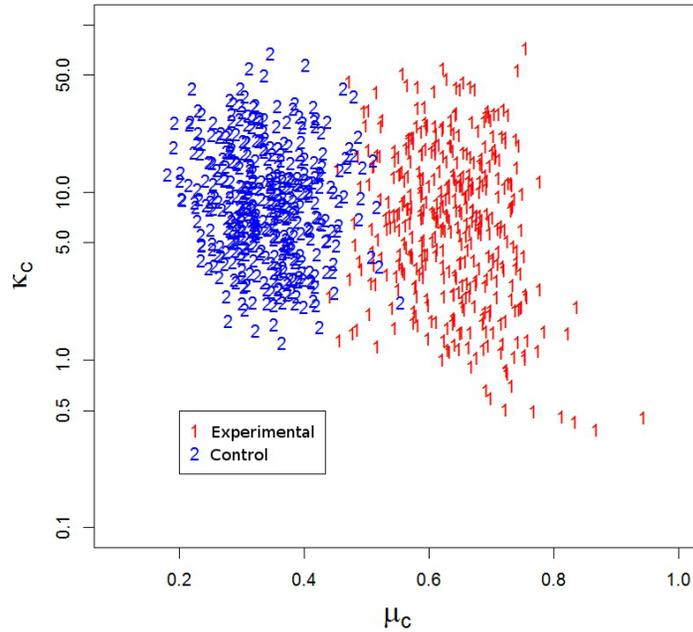
Figure 34. Mean posterior accuracies (μ_c) as a function of precision (k_c) for the experimental and control groups.

As is clear from Figure 34, the posterior distributions are clearly separated with the higher accuracy for the experimental group. Figure 35 shows the posterior means and their contrast. The mean accuracy for the experimental and the control groups were $\mu = 0.63$ and $\mu = 0.34$, respectively. As shown by the contrast, this difference is credible with a mean posterior difference of 0.29.
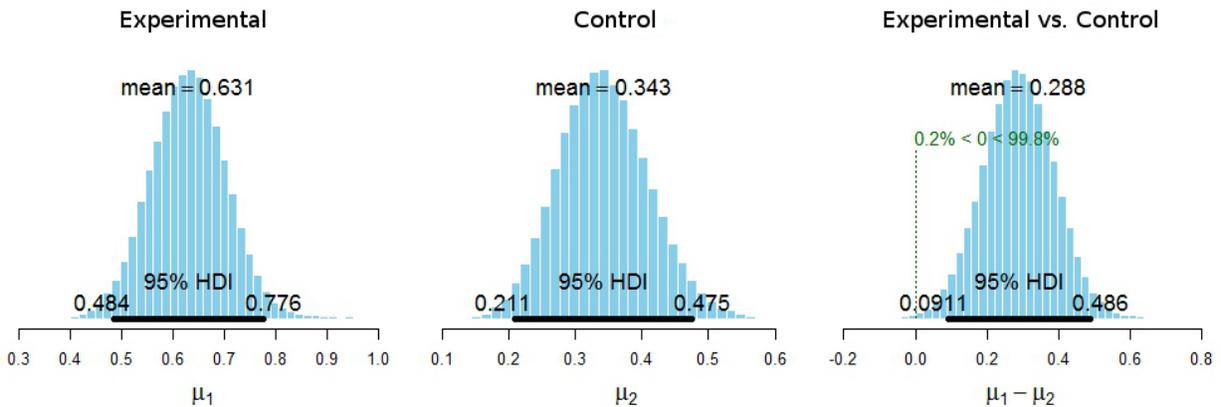
Figure 35. Mean posterior accuracies (left and middle) and contrast (right) for the experimental and control groups.

To follow up on the previous question that only required a yes or no answer, we asked a question that allowed us to get additional insight into pilot WSA: "How close to the precipitation cells did you get?" For this question, we provided four response alternatives, corresponding to four distance ranges: 20–25 nmi, 15–20 nmi, 10–15 nmi, and less than 10 nmi. For the control group, only pilots who answered yes to the previous question were included.

Figure 36 shows a summary of the question responses for the four distance categories. Although pilots from both groups provided responses in all four categories, the only correct answers were the responses for the 15–20 nmi category. Therefore, 41% ($n = 19$) of the pilots in the experimental group ($N = 46$) and 23% ($n = 3$) of the pilots from the control group ($N = 13$) correctly stated that they flew 15–20 nmi from the precipitation cells.
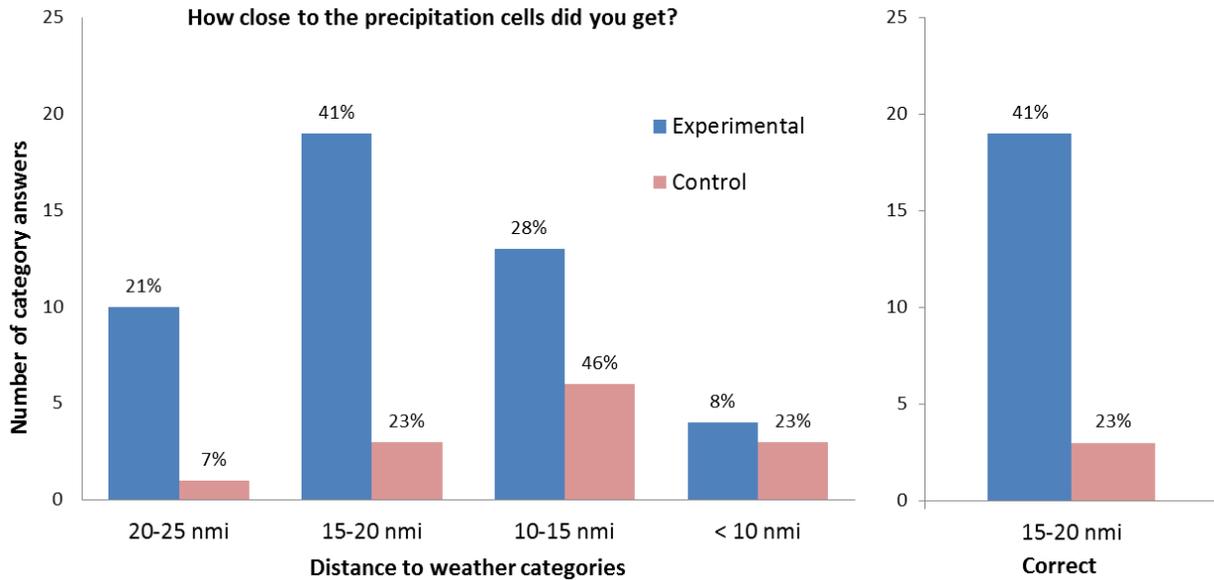
Figure 36. The number and percentage of answers to the four distance categories.

The proportion correct experimental responses is higher than the proportion control responses. However, using a Bayesian Beta-binomial model with MCMC sampling for contingency table analysis, we find that the control versus experimental difference is not credible. As can be seen in Figure 37, the posterior difference has a mean of -0.15 with the value zero included in the 95% HDI.
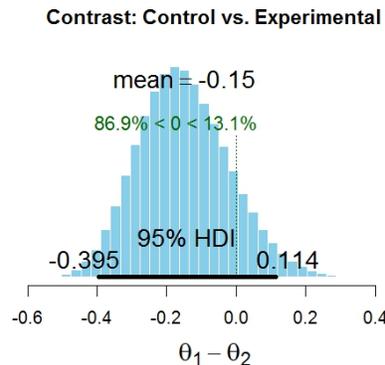
Figure 37. Posterior difference for the control versus experimental contrast.

Whereas the FAA and NOAA (1983) recommendation is to avoid hazardous storm cells by at least 20 statute miles, we were interested in pilots' subjective assessment of this distance. We

therefore asked the question: "The current FAA recommendation is to stay 20 nmi away from storms. Do you agree with this recommendation?" The outcome showed that 91% of all pilots agreed with the recommendation (96% for the experimental and 86% for the control). To probe this distance even further we asked: "If you would determine the recommendation for a distance to storms, what distance (in nmi) would you pick?" The majority of pilots selected the current recommended distance of 20 miles (71% for the experimental and 67% for the control), with only one pilot answering "less than 10 miles" (control), five pilots answering "10 miles" (control), four pilots answering "15 miles" (two from the experimental and control, respectively), and 21 pilots answering "30 miles or more" (12 pilots from the experimental and 9 pilots from the control). Consequently, there seems to be a consensus among pilots (90%) for keeping 20 or more miles away from hazardous storms.

Finally, we wanted to get pilot feedback (experimental group only) on the AR for the 20 nmi precipitation distance. We therefore asked the question: "During the scenario flight, the blue line reminder distance was set to 20 nmi. If you would use this implementation during actual flights, what nmi reminder distance would you set?" We provided four distance categories for the responses: (1) 20–25 nmi, (2) 15–20 nmi, (3) 10–15 nmi, and (4) less than 10 nmi. The outcome showed that the majority of pilots answered 15–20 nmi (58%), with 25% answering 20–25 nmi, 17% answering 10–15 nmi, and 0% answering less than 10 nmi. Therefore, 83% of the pilots agreed that the blue line AR for forecasted visibility areas should have a reminder distance between 15–25 nmi.

In summation, when we asked pilots what distance they would recommend from hazardous storms, 90% of the pilots agree on 20 or more miles. During the simulation, pilots in the experimental and control groups had an average closest point of approach to storm cells of 17.5 to 18 nmi, respectively. In a post-scenario questionnaire, pilots were asked if they flew closer than 20 nmi from precipitation cells. The number of correct answers to this question was credibly higher for the Experimental group (63%) compared to the Control group (33%), indicating a greater WSA for pilots in the Experimental group. When asked how close they flew to precipitation cells, 41% of pilots in the Experimental group and 23% of pilots in the Control group correctly stated that their closest point of approach was in the range of 15–20 nmi. This demonstrates a higher WSA for pilots in the experimental group compared to pilots in the control group. Evidently, the use of a weather display with a 20-nmi AR for precipitation cells increases pilot WSA by making pilots more cognizant of distances between the aircraft and storm cells.

3.2.4 The Use of AWOS Information During Flight

During the Alaska flight, pilots could tune in to three AWOS stations. These stations were located at JNU (Juneau International Airport), HNS (Haines airport), and SGY (Skagway airport). At the start of the scenario, JNU reported 10-mile visibility with 10,000 feet overcast. HNS reported 5-mile visibility and 6000 scattered, with a 5000 feet overcast. Finally, SGY reported 3-mile visibility and 5000 scattered, with a 4000 feet overcast. These weather reports were constant during the course of the scenario.

Similar to Simulation 1, we recorded each time pilots tuned in to an AWOS station. We then summarized these count values for each pilot and group (Experimental and Control). Figure 38 shows the total number of AWOS counts. The number of AWOS inquiries was 366 and 232 for the Experimental and Control groups, respectively. For the Experimental group, 43 of the 46

pilots (93%) of the pilots tuned in to one or more AWOS stations. For the Control group, 42 of the 49 pilots (86%) tuned in to one or more AWOS stations.
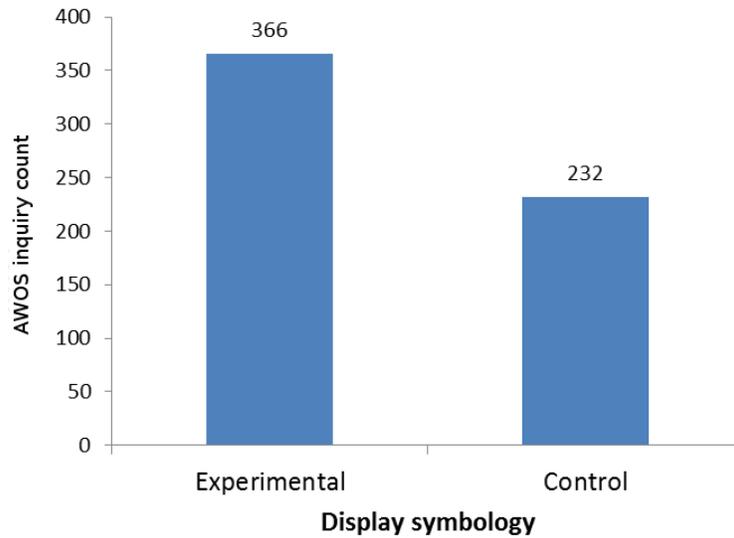
Figure 38. The total number of AWOS inquiries for the Experimental and Control groups.

The AWOS count is higher for the Experimental groups than the Control group. Using a Bayesian beta-binomial model with MCMC sampling, we find that the Control versus Experimental difference is credible (see Figure 39) with a mean posterior difference of 0.22 and the value 0 outside the 95% HDI.
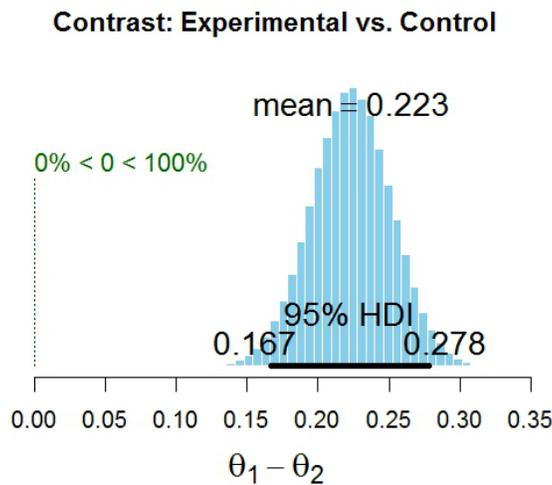
Figure 39. Posterior contrasts for the difference in AWOS usage between the experimental and control groups.

We also assessed the AWOS usage across the three AWOS stations. Figure 40 shows the Alaska AWOS stations (red dots) and their relation to the scenario start point (green dot).
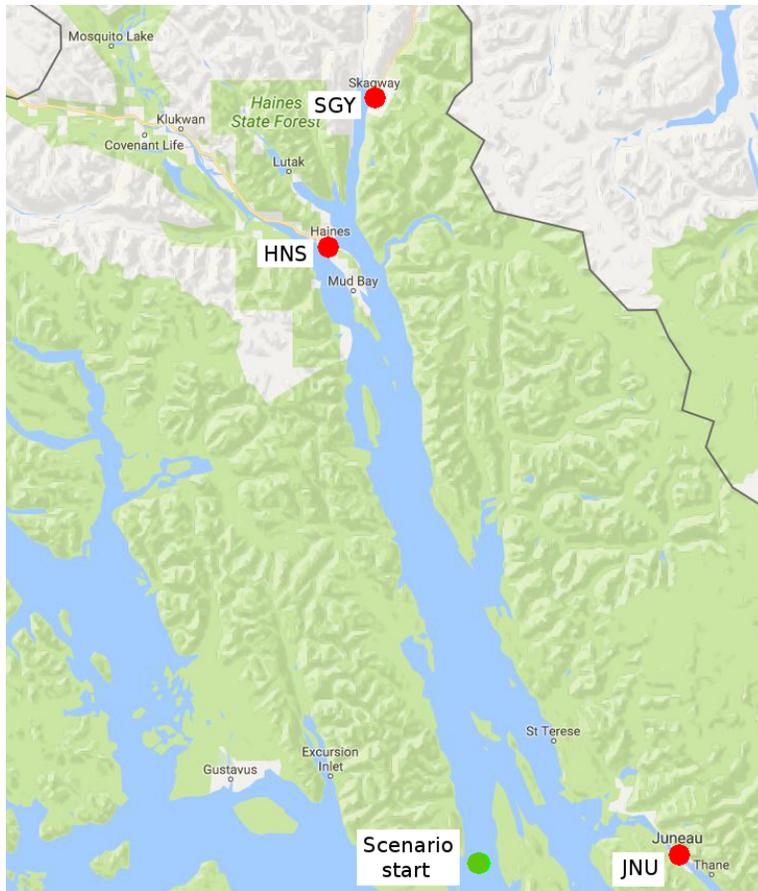
Figure 40. The three Alaska AWOS stations (red dots) at HNS, JNU, and SGY and their relation to the scenario start point (green dot).

Figure 41 shows the scenario time for each AWOS inquiry during the Alaska scenario. Of the three AWOS stations, the counts are highest for HNS (287), followed by SGY (160), and finally JNU (151). For the HNS station, pilots used the AWOS information during the entire simulation. For JNU, the majority of the AWOS inquiries were made at the beginning of the scenario and approximately 25 minutes into the scenario. For GSY, pilots used this information during the entire scenario but there is an elevated use between 25 and 30 minutes into the scenario.

The scenario times for these AWOS inquiries correspond to pilot decisions regarding deviations, turning around, or continuing flight to the destination airport. For example, the JNU station was mostly tuned in during the first five minutes of the scenario and at 25 minutes into the scenario. This likely reflects pilots that considered turning around within the first 5–6 min of the scenario. The elevated use at approximately 25 minutes is likely from pilots who were uncertain about the conditions at NHS and SGY, and therefore contemplated turning around and flying back to JNU. The SGY station has an elevated use after 20 minutes, which coincides with the reduced visibility as the flight progressed toward the destination airport. Here, the decision was either to continue to SGY, deviate to HNS, or turn around to JNU. The HNS station has a high usage during the entire scenario. This was expected because many pilots chose HNS as an alternate airport in case conditions deteriorated as the flight progressed toward SGY. Therefore, we would expect pilots to frequently check the HNS station during flight.

49
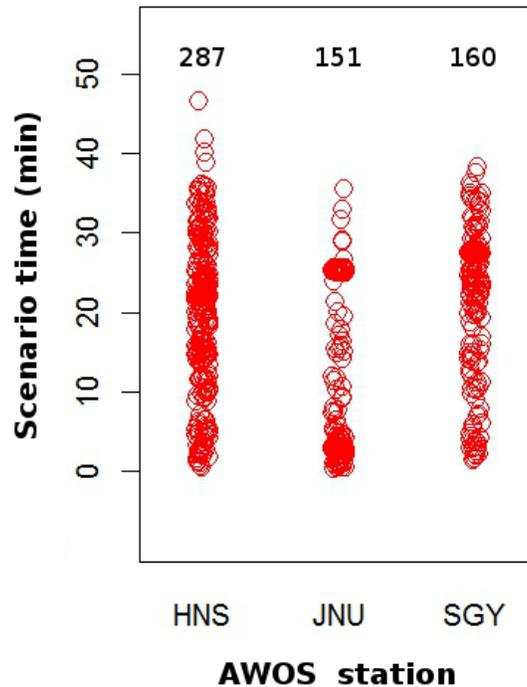
Figure 41. The scenario time for AWOS inquiries (red circles) at three AWOS stations. The numbers above the data columns represent the total count of AWOS inquiries.

In summation, AWOS information can increase pilot WSA and enhance pilot decision-making for decisions related to turning around, deviating, or continuing flight toward the destination airport. In the Alaska scenario, we found that the experimental group (N = 366 AWOS inquiries by 93% of the pilots) had a credibly higher AWOS usage then the control group (N = 232 AWOS inquiries by 86% of the pilots). The alternate airport at HNS received the highest count of AWOS inquiries, followed by the destination airport at SGY, and finally the AWOS station at JNU.

3.2.5 Out-the-Window Visibility Reports

During the Alaska scenario, we asked pilots to provide an estimate of the out-the-window visibility. These visibility estimates were initiated by ATC radio contacts: "November 520 Papa Sierra, say flight conditions" at 6, 12, and 19 minutes into the scenario.

In Simulation 1, pilots were flying in three different visibility zones while providing visibility estimates. In Simulation 2, however, pilots were still in the simulated 25 nmi visibility zone at the 6- and 12-minute visibility reporting intervals. This means that pilots provided two visibility estimates for the simulated 25 nmi visibility zone, and one visibility estimate for the 12 nmi visibility zone. Figure 42 shows the pilots' view from the cockpit at 5 minutes into the scenario.

Figure 42. View from the cockpit at 5 minutes into the Alaska scenario

From post-scenario questions, we know that pilots perceived the reductions in visibility during flight. We asked the question: "During flight, did you ever notice that the visibility was decreasing?" The outcome showed that 95% of the pilots were cognizant of the decreasing visibility ("Yes" answer). The remaining pilots deviated to JNU early in the scenario and therefore never entered different visibility zones ("No" answer).

In this section, we first provide an altitude analysis at the three reporting times. Next, we provide an analysis of pilot visibility estimates. Last, we provide an analysis of the relationship between pilot visibility estimates and aircraft altitude.

For this analysis, we used the aircraft altitude for each visibility estimate and the factors weather presentation (experimental and control) and reporting order (i.e., the 6-, 12-, and 19-minute marks labeled as First, Second, and Third, respectively). To analyze the data, we used the same model from Kruschke (2014) that was used for the altitude and visibility analysis in Simulation 1.

Figure 43 shows the aircraft altitude (red circles) for the experimental and control groups at the First, Second, and Third visibility reporting probes. As is shown by the altitude data in Figure 43, the central tendencies only show a small reduction in altitude across the reporting times. The altitudes are very similar for the experimental and control groups at the First and Second time marks, whereas they are lower at the Third time mark.
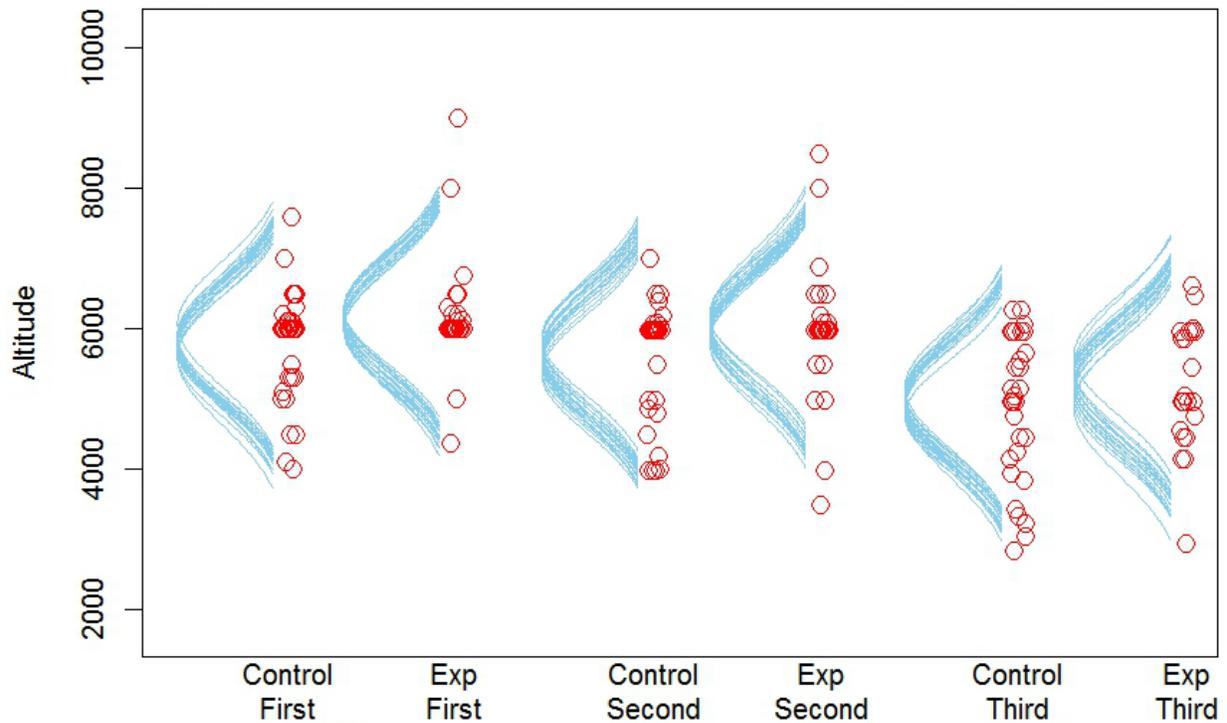
Figure 43. Individual aircraft altitudes (red circles) for the experimental and control groups at the First (6 min), Second (12 min), and Third (19 min) visibility estimation times.

Figure 44 shows the posterior contrasts for the difference in altitude between the experimental and control groups at the three reporting times. The comparison between the altitudes for the experimental and control groups at First (left) has a mode of 351 feet, with experimental having a higher altitude than control. However, because the value 0 is included in the 95% HDI, this difference is not credible. The comparison between the experimental and control groups at Second (middle) has a mode of 343 feet, implying that the experimental pilots were at higher altitudes than control pilots were. However, the 95% HDI for the posterior difference distribution includes the value zero, so this difference is not credible. Finally, the comparison of altitudes for the experimental and control groups at Third has a mode of 316 feet, implying that the experimental pilots were at higher altitudes than the control pilots were. Again, this difference is not credible because the 95% HDI includes the value zero.
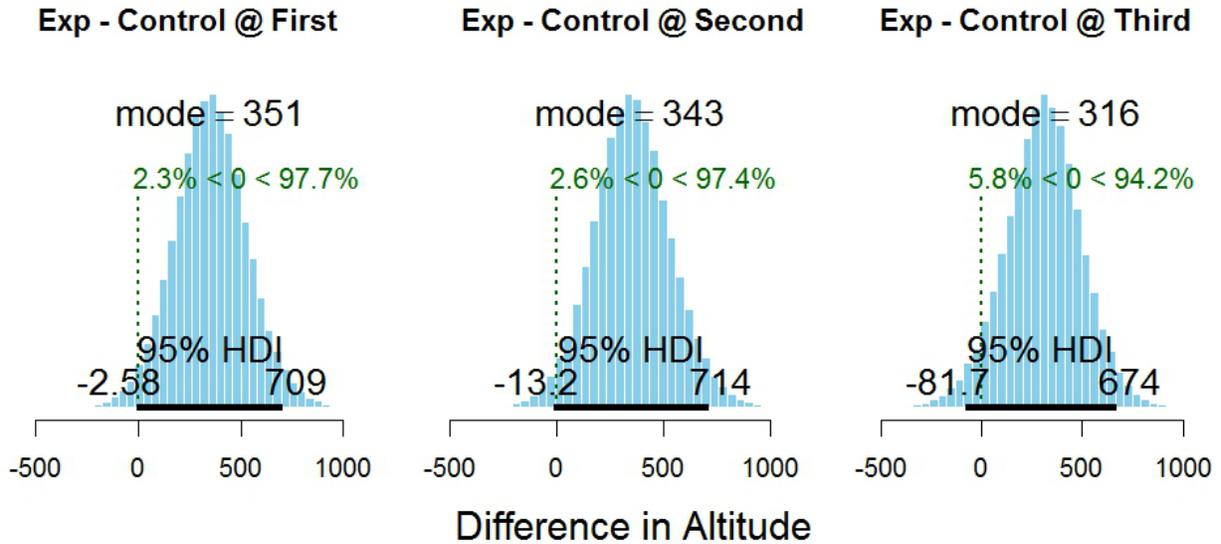
Figure 44. Posterior contrasts for the difference in aircraft altitudes between experimental and control groups at First (left), Second (middle), and Third (right).

Figure 45 shows pilot visibility estimates for the experimental and control groups at the simulated 12 nmi and 25 nmi visibility zones. The red circles in Figure 45 represent individual visibility estimates for the four combinations of weather presentation and visibility zone. The horizontal green lines represent a perfect correspondence between the reported and simulated visibilities. Similar to the result of Simulation 1, there is an underestimation of the reported visibility for both visibility zones.
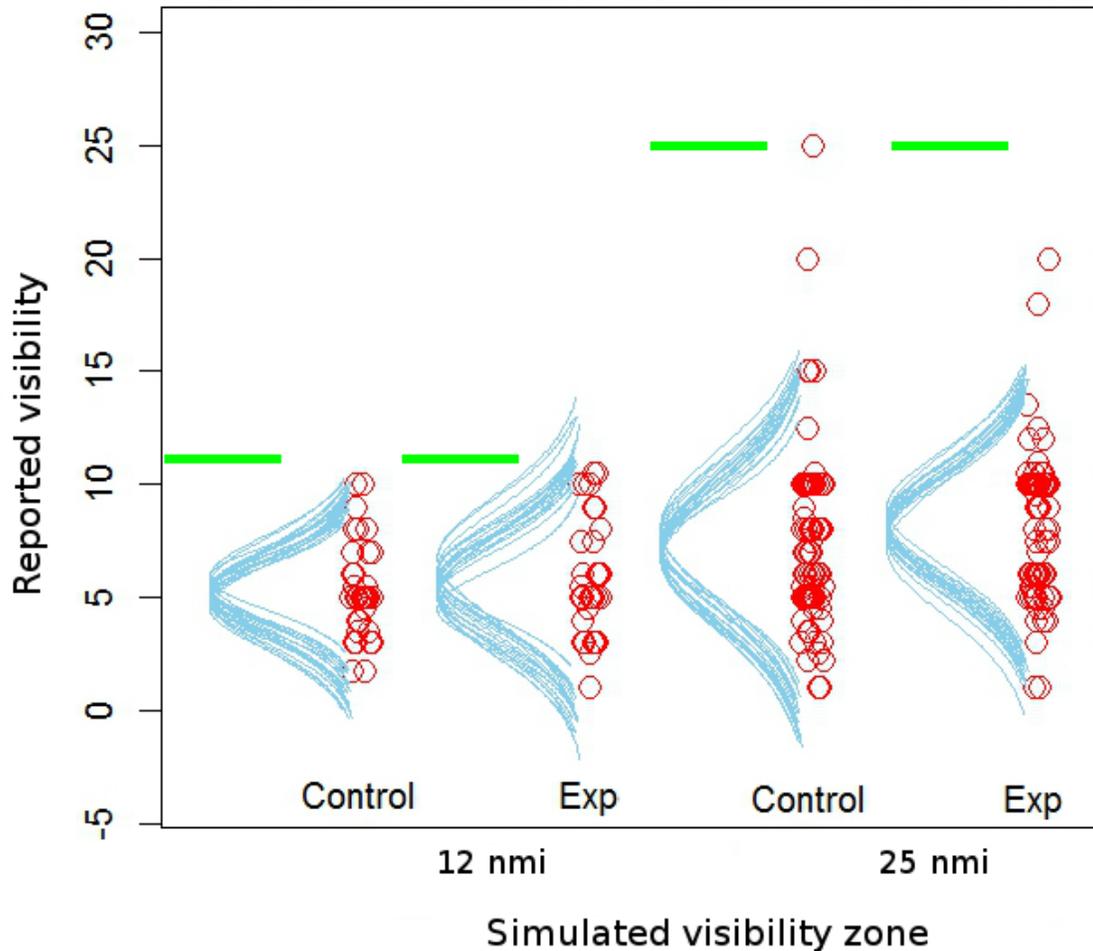
Figure 45. Individual visibility estimates (red circles) for the experimental and control groups at the 6-, 12-, and 19 nmi visibility zones. (Note: Pilots were still in the 25 nmi visibility zone for both the 6-minute and 12-minute visibility reporting.)

As shown in Figure 45 (left), the reported visibilities for the experimental and control groups at the 12 nmi visibility zone are very similar. All the reported visibility estimates are below the simulated visibility, denoted by the green line. The model analysis predicted a mean posterior visibility for the experimental and control groups of 5.7 (95% HDI from 4.9 to 6.8) and 5.2 (95% HDI from 4.6 to 5.8) miles, respectively.

For the 25 nmi visibility zone (right), there is an even greater underestimation of the simulated visibility with only one reported visibility estimate that coincided with the simulated visibility. The majority of the reported visibilities are 10 miles or below. The posterior mean for the experimental group was eight miles with a 95% HDI from 7.4 to 8.7. For the control group, the model predicted a mean of 7.3 miles with a 95% HDI from 6.6 to eight.

Figure 46 shows the posterior contrasts for a comparison of the visibility reports at the 12- and 25-nmi visibility zones and the comparison of reported visibilities between the experimental and control groups. As shown in Figure 46 (left), there is a credible difference between the visibility estimates for the 25- nmi and 12-nmi visibility zones with larger visibility estimates for the 25 nmi zone (mode of difference = 2.2). However, although the experimental group had slightly

larger estimates for both visibility zones, there is no credible difference in the visibility estimates between the experimental group and the control group (right). The mean posterior difference has a mode of 0.7 with the value zero included in the 95% HDI.
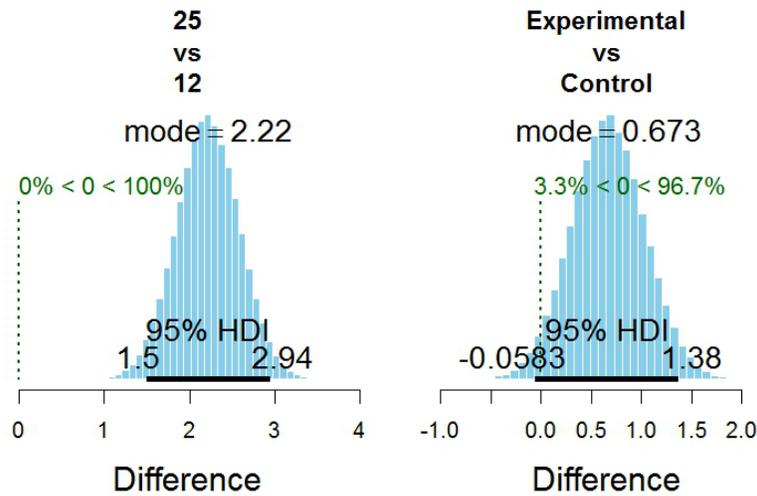
Figure 46. Posterior contrasts for comparisons of visibility estimates between the 25 nmi and 12 nmi visibility zones (left) and the difference between experimental and control groups (right).

Similar to the result from Simulation 1, we found no credible differences in pilot altitudes for this simulation. However, there are large variations among individual pilots in both altitude and visibility estimates. Therefore, similar to Simulation 1, we used the Kruschke (2014) regression model and the altitude (x) and corresponding visibility estimates (y) to assess the relationship between aircraft altitude and reported visibility.

We used the same prediction for this analysis as we used for the regression analysis in Simulation 1. If the aircraft altitude determined the reported visibility in our Alaska scenario, we expected to see a linear relationship between altitude (x) and reported visibility (y). Conversely, if there is no relationship between the aircraft altitude and reported visibility, we expected to see an estimated posterior slope of zero (indicating a lack of a relationship between x and y).

Figure 47 shows the outcome of the regression analysis on the altitude and visibility estimate data. We have aircraft altitude on the x-axis and reported visibility on the y-axis. The black circles are the data points (one altitude value per pilot visibility estimate). The lines represent regression lines and the three vertical distributions are superimposed t-distributed noise distributions. As shown by the regression lines in Figure 47, aircraft altitude does not seem to have a large effect on the reported visibilities. On the contrary, similar to Simulation 1, the reported visibilities for aircraft at the same altitude have a wide distribution of reported visibilities.
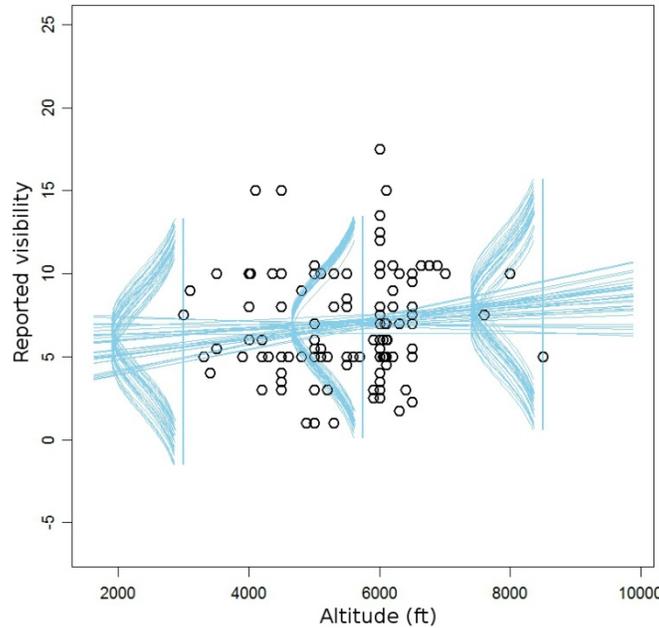
Figure 47. Regression lines and noise distributions for the prediction of reported visibility from aircraft altitude.

Figure 48 shows the mean posterior outcome for β_0 (intercept), β_1 (slope), and the scale parameter σ. The intercept has a mode of 5.08, which is the predicted value of y (visibility) when x (altitude) = 0. The non-credible slope (the value 0 is located within the 95% HDI) has a posterior mode of 0.0003. This means that as we increase altitude by one, visibility will increase by the value of β_1 (i.e., 0.0003). This outcome implies that aircraft altitude did not determine the reported simulation visibilities.
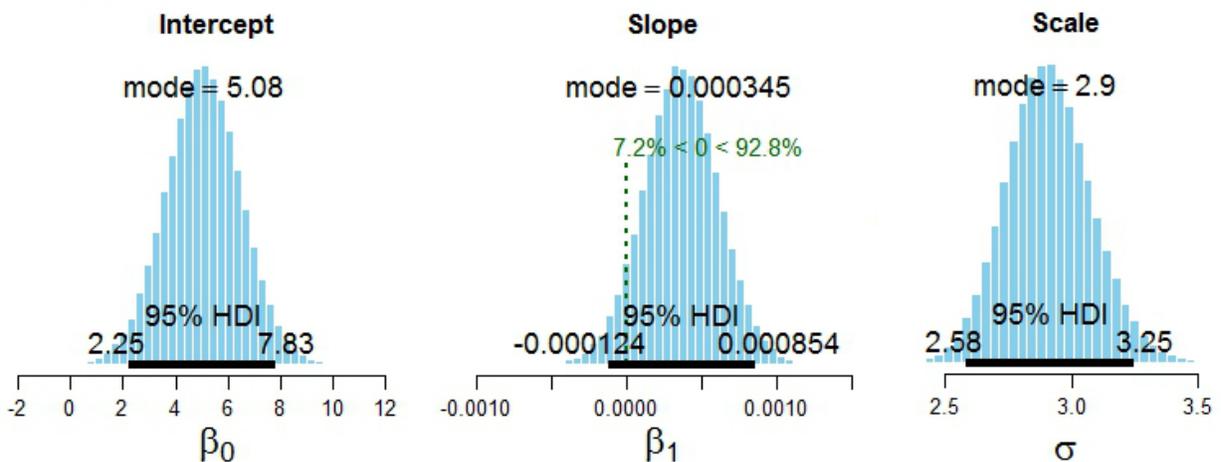
Figure 48. The posterior intercept (β_0), slope (β_1), and scale (σ) parameters from the robust linear regression analysis.

In summation, pilots provided visibility reports at three times during the simulation. However, pilots were still in the simulated 25 nmi visibility zone at the 6-minute and 12-minute visibility reporting, implying that pilots only provided visibility estimates for the simulated 25 nmi

visibility zone and the 12 nmi visibility zone. We correlated the reported visibilities with the simulated visibilities for the two visibility zones. We found that pilots greatly underestimated the simulated visibility in both visibility zones. By regressing the reported visibilities on aircraft altitude, we determined that the reported visibilities were not a function of aircraft altitude. The visibility estimate results are similar to the results from Simulation 1 and reveal a gap in pilot ability to determine out-the-window visibility accurately. The inaccurate visibility assessment speaks against a high WSA during flights in deteriorating visibility.

3.2.6 Decision-Making

In this section, we present data on pilot flight decisions in the six visibility zones (see Figure 29). Table 10 summarizes the pilot data by visibility zone of decision point for the Alaska simulation. (Note: Two pilots in the control group terminated the flight before deciding on a destination airport.) The table also lists the zones of visibility ranging from one nm to 25 nm and the total number of pilots that continued into each of the respective zones. For each visibility zone, the table shows the median age of the pilots, the number of IFR-rated pilots, number of non-rated pilots, median total flight hours, and the number of pilots in the experimental and control conditions. It also shows pilot decisions by indicating how many pilots decided to continue flight to HNS, SGY, and JNU.

Table 10. Pilot Flight Decisions at Six Simulated Visibility Zones.

Vis Zone	Total Pilots	Age	IFR rated	no rating	Flight Hours	Exp	Control	HNS	SGY	JNU
1 nmi	23	55	11	12	640	8	15	16	7	0
4 nmi	14	55	5	8	335	7	7	12	2	0
6 nmi	37	49	15	22	270	18	20	36	1	0
8 nmi	4	22	2	1	195.9	1	3	4	0	0
12 nmi	12	53	4	8	352	7	5	1	0	11
25 nmi	7	56	5	3	400	6	2	0	0	7

Note. In Table 1, the Age and Flight Hours are presented using the median (the median is the numerical value separating the upper half of a data sample from the lower half). For the 4nmi visibility range, one of the pilots did not report the rating.

Eight pilots made flight decisions in the 25 nmi visibility zone. Seven of these eight pilots decided to divert to JNU whereas one pilot decided to turn around.

For the 12 nmi visibility zone, 11 pilots decided to deviate to JNU and a single pilot decided to go to the alternate airport at HNS.

Only four pilots made a flight decision in the eight nmi visibility zone. All of these pilots made the decision to go to the alternate airport at HNS instead of continuing toward the destination airport at SGY.

For the six nmi visibility zone, 36 pilots made the decision to go to the alternate airport at HNS whereas one pilot (control) decided to continue flight toward the destination airport at SGY.

Fourteen pilots made a flight decision in the four nmi visibility zone. Twelve of these pilots decided to go to HNS whereas the remaining two pilots (Experimental and Control) decided to continue to the destination airport at SGY.

We have only summarized pilots that entered VFR zones (visibilities ranging from 25 nmi to four nmi) and made HNS, SGY, and JNU flight decisions under Visual Meteorological Condition (VMC). For the remaining one nmi visibility zone, however, pilots entered IMC conditions while under VFR. Twenty-three pilots entered and made their flight decisions in the one nmi visibility zone. Sixteen of these pilots decided to turn around and fly to the alternate airport at HNS whereas seven pilots decided to continue to SGY (1 pilot from the experimental group and 6 pilots from the control group). There were more pilots from the Control group (n = 15) than the Experimental group (n = 8) that entered the one nmi visibility zone before making a flight decision. As the posterior contrast in Figure 49 shows, the difference between the number of Control and Experimental group pilots entering the 1 nmi visibility zone is credible (mean posterior difference = 0.28).
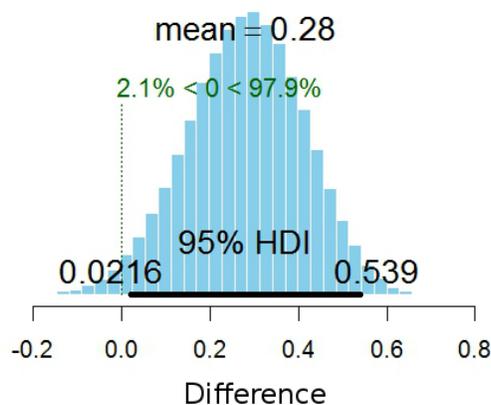
Figure 49. Posterior contrast for the comparison of flights into the one nmi visibility zone between the Control and Experimental groups.

In summation, during the scenario pilots encountered deteriorating visibility and made flight decisions to turn around to JNU, divert to the alternate airport at HNS, or continue flight toward the destination airport at SGY. For 75% of these flights, pilots were still in VMC when they made their decision. However, for the remaining 25% of the pilots, the decision to continue or divert came only after pilots had entered an area of IMC. In analyzing these VFR-into-IMC flights we found that there were credibly more pilots from the control group (n = 15) than the experimental group (n = 8) that entered the one nmi visibility zone. We attribute the lower

frequency of VFR-into-IMC flights as a direct result of the experimental group having access to an AR with forecasted visibility.

3.3 Summary of Findings for Simulation 2

1. Pilot/ATC communications showed the highest communication counts for AWOS communications, followed by communications related to ceiling-visibility-fog, turn around-divert, altitude change, and the lowest communication count for communications related to landing.

2. The use of a weather display with a 20 nmi AR for precipitation cells (experimental group) increased pilot WSA by making pilots more cognizant of distances between the aircraft and storm cells.

3. When asked if their closest point of approach to storm cells was less than 20 nmi, pilots in the experimental group (63%) had credibly more correct answers than the control group (33%). This indicates greater WSA for pilots in the experimental group.

4. When asked how close they flew to precipitation cells, 41% of pilots in the experimental group and 23% of pilots in the control group correctly stated that their closest point of approach was in the range of 15–20 nmi. Although this difference is not credible, it demonstrates a higher WSA for pilots in the experimental group compared to pilots in the control group.

5. The experimental group ($N = 366$ AWOS inquiries by 93% of the pilots) had a credibly higher AWOS usage than the control group (N = 232 AWOS inquiries by 86% of the pilots) during flight.

6. Pilots greatly underestimated simulated visibilities in the range of 12–25 nmi, and these results are not a function of aircraft altitude. This reveals a gap in pilots' ability to determine out-the-window visibility accurately and undermines a high WSA during flights in deteriorating visibility.

7. There were credibly more VFR-into-IMC flights (1 nmi visibility zone) for the control group (n = 15) than the experimental group (n = 8). We attribute the lower frequency of VFR-into-IMC flights as a direct result of the experimental group having access to an AR with forecasted visibility.

4. ADDITIONAL RESULTS

4.1 Mobile Device Proficiency Questionnaire

The Mobile Device Proficiency Questionnaire (MDPQ) is a tool developed and validated by Roque and Boot (2016) to assess mobile device proficiency in older adults (see Appendix D). The original questionnaire is comprised of 46 questions regarding operations on a smartphone or tablet device and rated on a 5-point scale (e.g., 1 = never tried; 2 = not at all; 3 = not very easily; 4 = somewhat easily; 5 = very easily). These questions are organized into eight subscales relating to: mobile device basics, communication, data and file storage, internet usage, calendaring, entertainment, privacy, and troubleshooting. A shorter version of the MDPQ, the MDPQ-16, reduces the questionnaire set to two questions per subscale.

Each participant completed the MDPQ-16 during the study. We analyzed all completed MDPQ-16 data regardless of whether or not the participants completed both simulations. Overall

proficiency scores for the MDPQ were calculated by averaging subscale questions and adding them to create one combined measure. A summary of MDPQ scores by age is provided in Table 11 and a regression analysis of all pilot proficiency scores is shown in Figure 50. The trend in mean proficiency scores for the Table 11 age bins (following the age bins used by FAA, 2016) is shown in Figure 51.

As shown in Figure 50, the regression lines have a negative slope meaning that the proficiency scores (y) decrease with an increasing age (x). The intercept has a mode of 41.8 (95% HDI from 40.2 to 43.7), which is the predicted proficiency score when age = zero. The credible slope has a posterior mode of -0.084 (95% HDI from -0.13 to -0.04) and a *SD* mode of 2.25 (95% HDI from 1.48 to 3.17). This means that as we increase age by one year, the predicted proficiency score decreases by 0.084. Therefore, overall, the result implies that older participants have lower mobile device proficiency scores than younger participants.

Table 11. Mean MDPQ Proficiency Scores by Age

Age bin	n	Mean Proficiency Score
15-19	2	39.5
20-24	9	38.5
25-29	6	39.3
30-34	5	36.8
35-39	8	39.0
40-44	7	39.8
45-49	7	38.7
50-54	6	37.1
55-59	17	36.8
60-64	10	37.1
65-69	8	31.8
70-74	7	33.3
75-79	4	32.0
80 and over	1	29.3
Total	97	
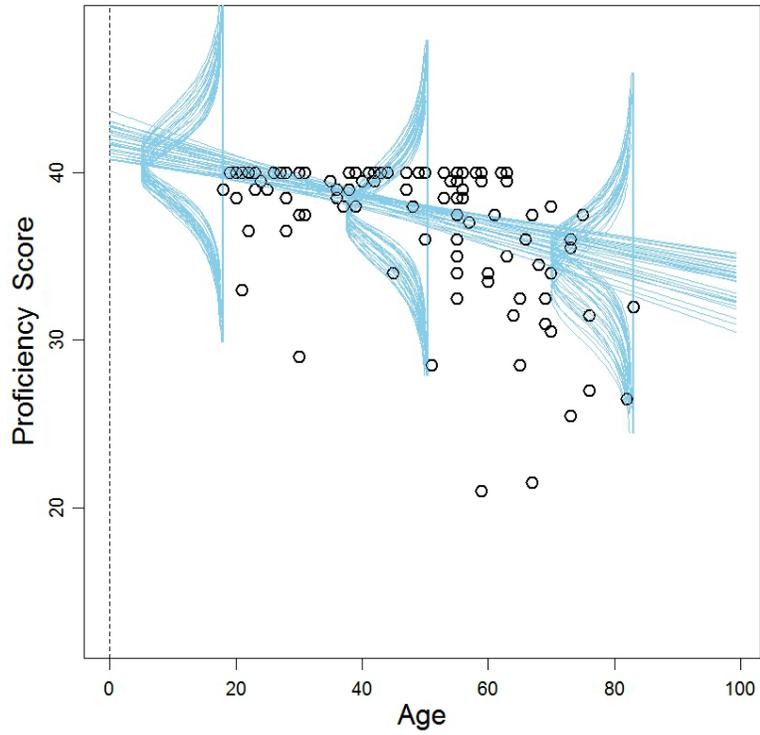
Figure 50. Linear regression of all mobile device proficiency questionnaire scores.
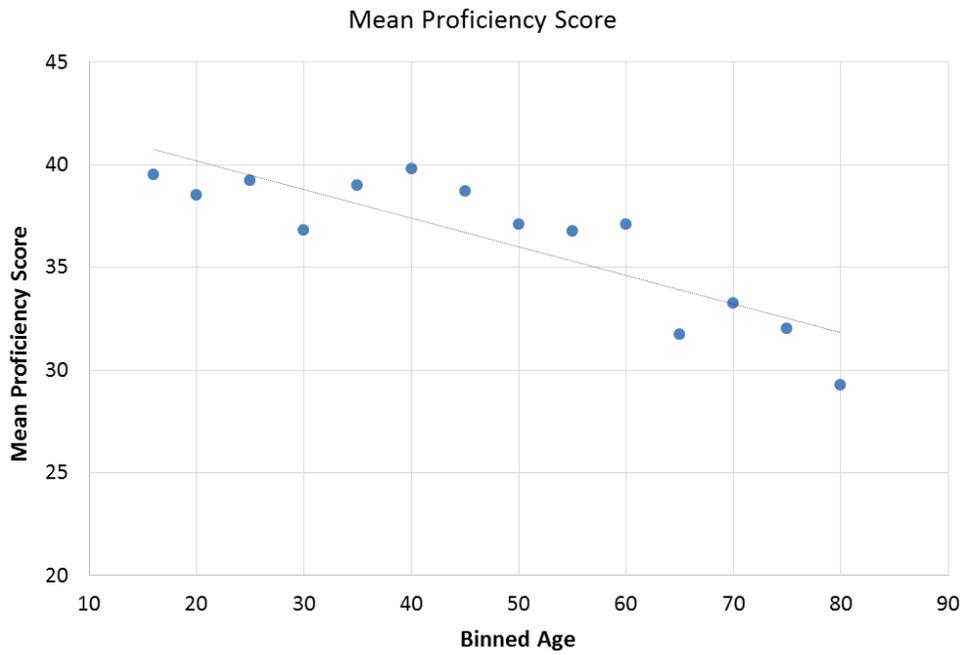
Figure 51. Mean mobile device proficiency questionnaire scores by age bins.

4.2 Weather Questionnaire

The Weather Questionnaire (see Appendix E) contains 20 questions selected from the private pilot written exam weather questions. Our long-term goal is to develop a standard set of questions that we use for all human-in-the-loop simulations. The difficulty is developing questions that tell us how effective a given question is in discriminating between pilots with "high" weather knowledge and pilots with "low" weather knowledge. Using the current data set, we used a hierarchical two-parameter logistic model for item response theory (IRT: http://doingbayesiandataanalysis.blogspot.com/search?q=IRT). In this analysis, each question (item) is modeled as producing 1 or 0 responses with a probability that depends on the question's difficulty and the pilot's knowledge. The model allows comparisons of question difficulties, question discriminations, or pilot abilities. However, the Bayesian IRT analysis did not provide enough information as the test requires more data to be meaningful. Therefore, we only provide summary statistics (see Table 12) and a regression analysis for the effect of age on pilot weather knowledge. Table 12 summarizes the median and range of the overall weather questionnaire responses for both Simulation 1 (PA) and Simulation 2 (AK) by condition and group. Questionnaire performance appears to be homogeneous across groups with similar median and range results. The relatively low median score with the large variability in score range indicates needed improvement in terms of weather-related training and education.

Table 12. Summary Data for Weather Questionnaire

Condition	Weather Question Performance			
	Group	n	Median	Range
PA	D1	47	57%	20%-85%
	D3	47	65%	30%-90%
	All	94	60%	20%-90%
AK	Weather App	48	65%	20%-95%
	No Weather App	50	53%	30%-85%
	All	98	60%	20%-95%
All	All	101	60%	20%-95%

Figure 52 shows the mean score (correctly answered questions) for each of the different age bins. For the averaged data set, there seems to be a trend in the data in which the proportion correct decreases with increasing age.
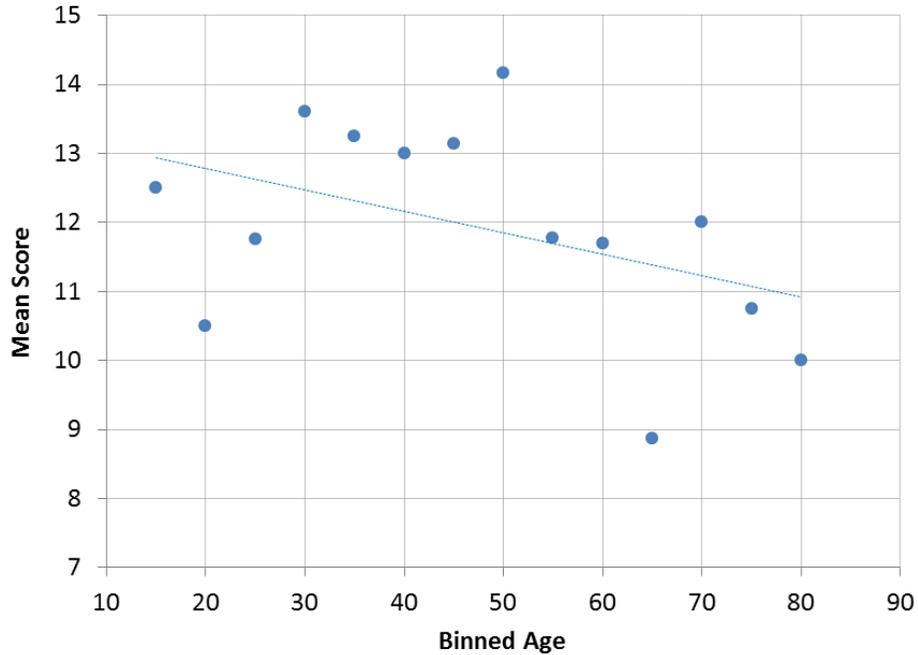
Figure 52. Mean weather question score as a function of age bin.

However, using all the pilot weather questionnaire scores in a linear regression revealed that the negative slope is not credible. The outcome showed a posterior intercept with a mode of 13 (95% HDI from 11.1 to 15) and a slope with a mode of -0.02 (95% HDI from -0.06 to 0.01) where the value 0 was included in the 95% HDI. There is also a large dispersion in the data, yielding a posterior *SD* with a mode of 3.12 (95% HDI from 2.67 to 3.64). Therefore, there is no credible decline in weather scores with an increasing age.
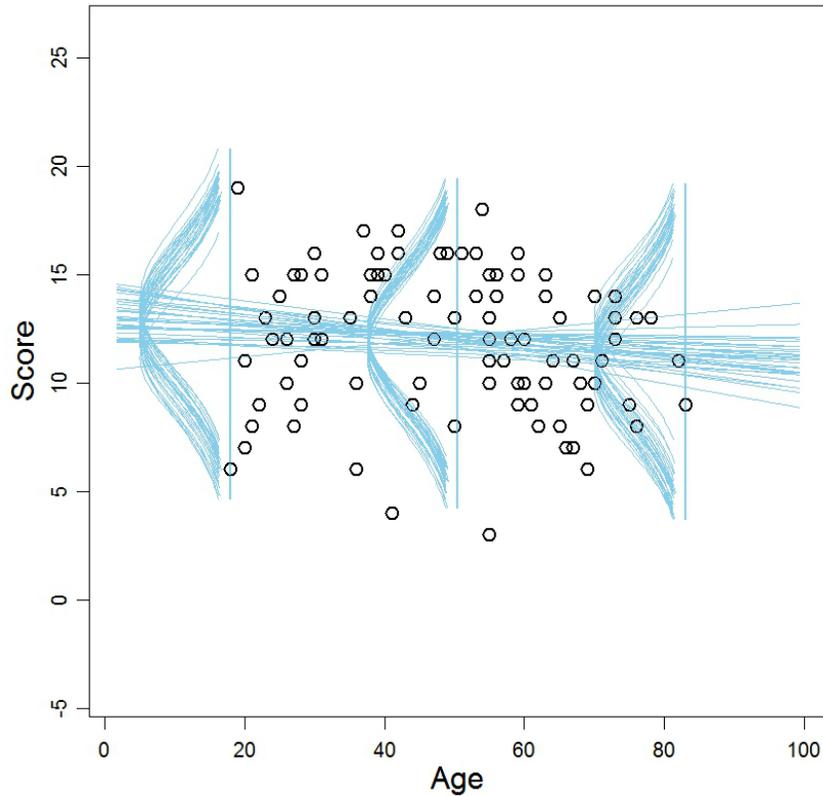
Figure 53. Linear regression of all pilot weather questionnaire scores and ages.

In addition to answering weather questions, pilots also rated their confidence that their answers were correct (scale from 30% to 100%). Figure 54 shows the mean confidence ratings as a function of proportion correct for the 20 weather questions. This plot tells us how well "calibrated" pilots are with regard to their weather knowledge. Ideally, the 20 data points would line up along the diagonal black line. This would imply that pilots are very cognizant of their weather knowledge. For example, if pilots' average confidence rating for a given question is 0.6, then the average proportion correct should also be 0.6. In Figure 54, we see the opposite as 75% of the data points are below the black diagonal line. This implies that pilots are overconfident; they think they know more about weather than they actually do. If the data points would instead be above the black diagonal line, pilots would be "under confident"; in fact, they would have more weather knowledge than they think they have.
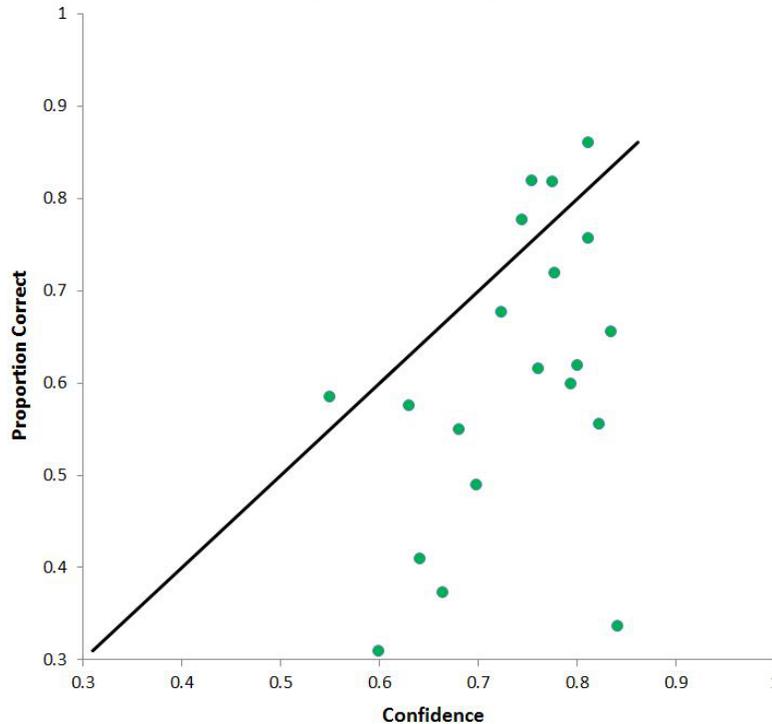
Figure 54. Mean confidence ratings as a function of proportion correct answers for each of the 20 weather questions.

5. DISCUSSION

In this study, we examined pilot use of weather information during flights in deteriorating visibility. We measured key variables related to pilot situation assessment, WSA, and pilot decision-making. Here, we discuss the study and relate our findings to previous research on GA flights in deteriorating weather. First, we discuss the effect of symbol salience enhancements on pilot use of METAR information. Second, we discuss pilot use of "blue line" ARs for precipitation and visibility forecasts and how the ARs affected pilot WSA and decision-making. Third, we discuss pilot situation assessments and the difficulty in making estimates of distances to storms and out-the-window visibility.

Previous GA research has uncovered a range of instances in which pilots failed to detect, integrate, and use graphical METAR information to enhance situational assessments and flight decisions (Johnson, Wiegmann, & Wickens, 2006; Coyne, Baldwin, & Latorella, 2005). Ahlstrom and Suss (2014) and Ahlstrom et al. (2015b) found that METAR symbology (i.e., symbol shape and color) affected whether pilots noticed and used the METAR information during flight. Differences in the presentation symbology yielded variations in how often pilots noticed and used METAR information, with the recognition of METAR updates (i.e., VFR-to-IFR) ranging from 25% to 62%. Pilot performance was directly related to METAR symbol salience, with a higher salience difference for the METAR symbol changes that yielded 62% detection accuracy compared to the METAR symbol changes that yielded 25% detection accuracy. Pilots who failed to detect and use updated METAR information were more likely to continue their VFR flight toward IMC conditions.

In this study, we subjected our METAR symbols to a frequency-tuned salience analysis (using an algorithm by Achanta, Hemami, Estrada, and Süsstrunk, 2009) and enhanced our METAR symbols by increasing the symbol salience. Our results show no difference in pilot use of METAR symbols (i.e., triangles and circles) with VFR-to-IFR detection rates of 52% and 62%. However, even though the detection rate for METAR triangles in this study is higher than what was found by Ahlstrom and Suss (2014), the effect of increasing the METAR symbol salience on change-detection performance is rather small. Using a Bayesian meta-analysis of available METAR change-detection data, we determined that the mean overall effect on performance from an increase in METAR symbol salience is only approximately 12%. Therefore, in light of the small performance gains, we believe that METAR color changes (i.e., salience differences) are not enough to guarantee an unfailing detection by pilots. The main reason is the priority of pilots' divided visual attention and visual scan during single-pilot operations. Pilots are often restricted to a quick extraction of display information from one glance to the next. This is a recipe for change blindness, causing a failure by pilots to encode and compare the weather display information from only a few visual samples (Durlach, 2004; Rensink, O'Regan, & Clark, 2000). With this in mind, we propose that METAR weather display updates be coupled with other notification modalities such as vibrotactile cues (Ahlstrom et al., 2015b; Ferris & Sarter, 2011; Hameed, Ferris, Jayaraman, & Sarter, 2009).

The coupling of display information and vibrotactile notifications could also be extended to the AR for precipitation displays and visibility forecasts. Previous research has found that pilots fly too close to hazardous precipitation areas and they have great difficulty in making estimates of out-the-window visibility (Ahlstrom et al., 2015a, 2015b; Coyne, Baldwin, & Latorella, 2008; Goh & Wiegmann, 2001; Wiegmann, Goh & O'Hare, 2002). In this study, we used a blue line display to notify pilots that they were 20 nmi away from 30 dBZ precipitation cells or 20 nmi away from forecasted areas of 3 nmi or less visibility. The result shows clear benefits of the AR display with a credibly higher WSA for the experimental group compared to the control group (no AR). This was expressed as a clearer pilot understanding of the closest point of approach to hazardous precipitation cells and credibly less entries into one nmi visibility zones (i.e., IMC). When asked about the safety distance, 91% of all study pilots agreed with the current recommendation to avoid hazardous storm cells by at least 20 statute miles. However, when asked if they flew closer than 20 miles during the simulation, only between 33% (control) and 63% (experimental) of pilots provided a correct answer. Consequently, there still is a gap between pilots' intent to stay 20 miles away from hazardous storms and their ability to accurately estimate distances.

We also found that on average pilots underestimate the out-the-window visibility for a range of simulated visibilities. There was no consensus among pilots in their visibility reports, expressed as large dispersions (i.e., variability) of the estimates. Furthermore, half of the Alaska scenario pilots who entered the four nmi visibility zone were equipped with the AR for visibility forecasts. The fact that pilots have difficulty in making out-the-window visibility estimates is problematic as previous research has uncovered that pilots use horizontal visibility and cloud concentration information to assess VFR conditions (Wiggins & O'Hare, 2003). In Simulation 2, only 17% of the pilots could correctly report their lowest visibility encounter during flight. Furthermore, a majority of these correct reports (76%) were from pilots who entered approximately one nmi visibility zones (i.e., IMC). This reveals a gap in pilot situational assessments and undermines a high WSA during flights in deteriorating visibility. The failure to assess current visibility conditions increases the odds of VFR-into-IMC flights. Therefore, we

believe a coupling of vibrotactile notifications with the AR for precipitation and visibility thresholds could enhance pilot WSA even further by providing explicit and timely notifications of distances to hazardous conditions along the route of flight.

References

Achanta, R., Hemami, S., Estrada, F., & Süsstrunk, S. (2009). *Frequency-tuned Salient Region Detection*. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009), 1597 - 1604.

Ahlstrom, U. (2015). Experimental Evaluation of the AIRWOLF Weather Advisory Tool for En Route Air Traffic Controllers. *Aviation Psychology and Applied Human Factors, 5*(1), 18-35. https://doi.org/10.1027/2192-0923/a000070

Ahlstrom, U. (2015). Weather Display Symbology Affects Pilot Behavior and Decision-Making. *International Journal of Industrial Ergonomics, 50,* 73-96. https://doi.org/10.1016/j.ergon.2015.09.008

Ahlstrom, U., Caddigan, E., Schulz, K., Ohneiser, O., Bastholm, R., & Dworsky, M. (2015a). *Initial Assessment of Portable Weather Presentations for General Aviation Pilots* (DOT/FAA/TC-15/42). Atlantic City International Airport, NJ: FAA William Hughes Technical Center.

Ahlstrom, U., Caddigan, E., Schulz, K., Ohneiser, O., Bastholm, R., & Dworsky, M. (2015b). *The Effect of Weather State-change Notifications on General Aviation Pilots' Behavior, Cognitive Engagement, and Weather Situation Awareness* (DOT/FAA/TC-15/64). Atlantic City International Airport, NJ: FAA William Hughes Technical Center.

Ahlstrom, U., & Dworsky, M. (2012). *Effects of Weather Presentation Symbology on General Aviation Pilot Behavior, Workload, and Visual Scanning* (DOT/FAA/TC-12/55). Atlantic City International Airport, NJ: FAA William Hughes Technical Center.

Ahlstrom, U., & Jaggard, E. (2010). Automatic Identification of Risky Weather Objects in Line of Flight (AIRWOLF). *Transportation Research Part C: Emerging Technologies*, *18*(2), 187-192.

Ahlstrom, U., & Racine, N. (2019). *Symbol Salience Augments Change-Detection Performance in Cockpit Weather Displays.* (DOT/FAA/TC-19/31). Atlantic City International Airport, NJ: FAA William Hughes Technical Center.

Ahlstrom, U., & Suss, J. (2014). *Now You See Me, Now You Don't: Change Blindness in Pilot Perception of Weather Symbology* (DOT/FAA/TC-14/16). Atlantic City International Airport, NJ: FAA William Hughes Technical Center.

Barnabas, A. F., & Sunday, A. O. (2014). Comparison of allocation procedures in a stratified random sampling of skewed populations under different distributions and sample sizes. *International Journal of Innovative Science, Engineering & Technology, 1*(8), 218-225. ISSN 2348 – 7968.

Batt, R., & O'Hare, D. (2005). Pilot behaviors in the face of adverse weather: A new look at an old problem. *Aviation, Space, and Environmental Medicine, 76*(6), 552-559. ISBN 1 921092 09 2.

Beringer, D., & Ball, J. (2004). *The effects of NEXRAD graphical data resolution and direct weather viewing on pilots' judgments of weather severity and their willingness to continue*

flight (DOT/FAA/AM-04/5). Oklahoma City, OK: Civil Aerospace Medical Institute, Federal Aviation Administration.

Burgess, M. A., & Thomas, R. P. (2004). *The effect of NEXRAD image looping and national convective weather forecast product on pilot decision making in the use of a cockpit weather information display* (Tech. Rep. NASA/CR-2004-212655). Hampton, VA: Langley Research Center.

Coyne, J. T., Baldwin, C. L., & Latorella, K. A. (2005). Influence of graphical METARs on pilot's weather judgment. In *Proceedings of the Human Factors and Ergonomics Society 49th annual meeting*. Santa Monica, CA: Human Factors and Ergonomics Society, 131–135. doi: 10.1177/154193120504900129

Coyne, J. T., Baldwin, C. L., & Latorella, K. A. (2008). Pilot Weather Assessment: Implications for Visual Flight Rules Flight Into Instrument Meteorological Conditions. *International Journal of Aviation Psychology, 18*(2), 153 – 166. https://doi.org/10.1080/10508410801926756

Durlach, P. J. (2004). Army digital systems and vulnerability to change blindness. In H. Kwon, N. M. Nasrabadi, W. Lee, P. D. Gader, & J. N. Wilson (Eds.), *Proceedings of the Twenty-Fourth Army Science Conference* (Accession No. ADM001736). Redstone Arsenal, AL: Army Missile Research, Development and Engineering Lab.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors, 37*, 65–84. https://doi.org/10.1518/001872095779049499

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of Convenience Sampling and Purposive Sampling. *American Journal of Theoretical and Applied Statistics, 5*(1), 1-4. doi: 10.11648/j.ajtas.20160501.11.

FAA (2016). U.S. Civil Airmen Statistics. https://www.faa.gov/data_research/aviation_data_statistics/civil_airmen_statistics/

FAA, & NOAA. (2010). *Aviation weather services AC 00-45*. Oklahoma City, OK: N.O.A.A and F.A.A.

Ferris, T. and Sarter, N. (2011). Continuously-informing vibrotactile displays in support of attention management and multitasking in Anesthesiology. *Human Factors, 53*(6), 600-611. doi: 10.1177/0018720811425043

Goh, J., & Wiegmann, D. A. (2001). Visual flight rules flight into instrument meteorological conditions: An empirical investigation of the possible causes. *International Journal of Aviation Psychology,11*, 359–379. https://doi.org/10.1518/0018720024497871

Hameed, S., Ferris, T., Jayaraman, S., and Sarter, N. (2009). Using informative peripheral visual and tactile cues to support task and interruption management. *Human Factors, 51*(2), 126-135. https://doi.org/10.1177/0018720809336434

Hardy, D. J., & Parasuraman, R. (1997). Cognition and flight performance in older pilots. *Journal of Experimental Psychology: Applied, 3*(4), 313-348.

Hunter, D. R., Martinussen, M., Wiggins, M., & O'Hare, D. (2011). Situational and personal characteristics associated with adverse weather encounters by pilots. *Accident Analysis and Prevention, 43*, 176–186. doi:10.1016/j.aap.2010.08.007

Johnson, N., Wiegmann, D., & Wickens, C. (2006). Effects of advanced cockpit displays on general aviation pilots' decision to continue visual flight rules flight into instrument meteorological conditions. In *Proceedings of the Human Factors and Ergonomics Society 50th annual meeting*. Santa Monica, CA: Human Factors and Ergonomics Society, 30-34. https://doi.org/10.1177/154193120605000107

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis: a Tutorial with R and BUGS*. Academic Press/Elsevier, Burlington, MA. http://dx.doi.org/10.1016/jtics201005001.

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: a Tutorial with R, JAGS, and Stan, 2nd edition*. Academic Press/Elsevier, ISBN 9780124058880.

Kruschke, J. K., & Liddell, T. M. (2016). *The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Planning from a Bayesian Perspective*. https://doi.org/10.2139/ssrn.2606016

Plummer, M., 2003. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vienna, Austria. Retrieved from http://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf.

Plummer, M., 2011. RJAGS: Bayesian Graphical Models Using MCMC. *R Package Version 3-5* [Computer software]. Retrieved from. http://CRAN.R-project.org/ package/rjags.

R Development Core Team, 2011. R: a Language and Environment for Statistical Computing [Computer software manual]. *R Foundation for Statistical Computing*, Vienna. Retrieved from http://www.R-project.org.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (2000). On the failure to detect changes in scenes across brief interruptions. *Visual Cognition, 7*(1/2/3), 127-145. doi: 10.1080/135062800394720

Roque, N. A., & Boot, W. R. (2016). A New Tool for Assessing Mobile Device Proficiency in Older Adults: The Mobile Device Proficiency Questionnaire. *Journal of Applied Gerontology*, 1-26. https://doi.org/10.1177/0733464816642582

Wiegmann, D. A., Goh, J., & O'Hare, D. (2002). The role of situation assessment and flight experience in pilots' decisions to continue visual flight rules flight into adverse weather. *Human Factors, 44*, 189–197. https://doi.org/10.1518/0018720024497871

Wiggins, M. W., & O'Hare, D. O. (2003). Expert and novice pilot perceptions of static in-flight images of weather. *The International Journal of Aviation Psychology, 13*(2), 173-187. doi: 10.1207/S15327108IJAP1302_05

Wiggins, M., & O'Hare, D. (2003). Weatherwise: Evaluation of a cue-based training approach for the recognition of deteriorating weather conditions during flight. *Human Factors, 45*, 337–345. https://doi.org/10.1518/hfes.45.2.337.27246

Wilson, D. R., & Sloan, T. A. (2003). VFR flight into IMC: Reducing the hazard. *The Journal of Aviation/Aerospace Education & Research, 13*(1). Retrieved from http://commons.erau.edu/jaaer/vol13/iss1/9.

Wu, S-C., Gooding, C. L., Shelley, A. E., Duong, C. G., & Johnson, W. W. (2012). Pilot convective weather decision making in en route airspace. In *Proceedings of the ICAS 2012—*

28th Congress of the International Council of the Aeronautical Sciences. Retrieved from http://www.icas.org/ICAS_ARCHIVE/ICAS2012/PAPERS/904.PDF

Acronyms

AR	Active Reminder
ASOS	Automated Surface Observing System
ATC	Air Traffic Control
AWOS	Automated Weather Observing System
EMI	Westminster VOR
ETX	East Texas VOR
FAA	Federal Aviation Administration
GA	General Aviation
HDI	High Density Interval
HSI	Horizontal Situation Indicator
IFR	Instrument Flight Rules
IMC	Instrument Meteorological Conditions
IRT	Item Response Theory
KABE	Allentown, Pennsylvania
KBWI	Baltimore Washington International, MD
KDCA	Washington National, DC
KESN	Easton/Newman, MD
KHGS	Hagerstown, MD
KIAD	Washington, DC/Dulles, VA
KJST	Johnstown, PA
KMRB	Martinsburg, West Virginia
LRP	Lancaster VOR
MCMC	Markov Chain Monte Carlo
MDPQ	Mobile Device Proficiency Questionnaire
METAR	aviation routine weather report
NEXRAD	Next Generation Radar
PAGY	Skagway Airport
PAJN	Juneau International Airport
PTT	Push-to-Talk
ROPE	Region of Practical Equivalence
SAGAT	Situation Awareness Global Assessment Technique
SME	Subject Matter Expert
VFR	Visual Flight Rules
VHF	Very High Frequency
VMC	Visual Meteorological Conditions
VOR	Very High Frequency Omnidirectional Radio Range
WSA	Weather Situation Awareness
WTIC	Weather Technology in the Cockpit

Appendix A

Informed Consent Statement

I, _____, understand that this cockpit study, entitled "Human Factors Laboratory Assessment of Meteorological (MET) Presentations IV" is sponsored by the Federal Aviation Administration (FAA) and is being directed by Ulf Ahlstrom.

Nature and Purpose:

I volunteered as a participant in this study that encompasses two part-task experiments and two cockpit simulations. The primary purpose of the experiments and the cockpit simulations is to improve weather presentations for the cockpit. During the experiments, participants will evaluate sets of static images displayed on a computer monitor. During the simulation flights, participants will fly two different single-engine general aviation (GA) simulators in deteriorating weather conditions.

Research Procedures:

Ninety private GA pilots will participate during a half-day (4 hours) session that covers two simulation flights and two computer experiments. The participants will be engaged from 8:00 AM to 12:00 PM (or from 12:00 PM to 16:00 PM) with short rest breaks. All the participants will conduct the simulator flight before performing the computer experiments.

The first part of the session will encompass a self-paced PowerPoint briefing to review project objectives and participant rights and responsibilities. This briefing will also include initial familiarization training on the cockpit simulators and the portable weather application. The participant will first complete a practice flight scenario. After the practice scenario, the participant will fly a designated route during a simulator flight (approximate duration: 40 minutes). During the simulator flight, an automated data-collection system will record cockpit system operations and generate a set of standard simulation measures including ATC communications. I understand that the recordings will include continuous audio and video recording of my actions in the cockpit for the duration of the flight.

After the simulation flight, the participants will complete a questionnaire to report their overall workload, weather situation awareness, and weather presentation usability and provide an assessment of the cockpit system and test condition.

After completing the questionnaire, participants will be briefed on the computer experiment and thereafter conduct a training session. After the training session participants will complete the experimental task, which is divided into blocks. During these blocks, an automated data collection system records each participant response.

Anonymity and Confidentiality:

The information that I provide as a participant is strictly confidential and I shall remain anonymous. I understand that no Personally Identifiable Information [PII] will be disclosed or released, except as may be required by statute. I understand that situations when PII may be disclosed are discussed in detail in FAA Order 1280.1B "Protecting Personally Identifiable Information [PII]."

Benefits:

I understand that the only benefit to me is that I will be able to provide the researchers with valuable feedback and insight into weather presentation symbology. My data will help the FAA to establish human factors guidelines for weather displays and assess if there is a need to standardize the symbology for enhanced weather information.

Participant Responsibilities:

I am aware that to participate in this study I must be a private GA pilot. I am also aware that I am not allowed to participate if I have a personal and/or familial history of epilepsy.

I will (a) fly the designated routes in the two cockpit simulations, (b) perform the two part-task experiments, and (c) answer questions asked during the study to the best of my abilities. I will not discuss the content of the experiments or the cockpit simulations with other potential participants until the study is completed.

Participant Assurances:

I understand that my participation in this study is voluntary and I can withdraw at any time without penalty. I also understand that the researchers in this study may terminate my participation if they believe this to be in my best interest. I understand that if new findings develop during the course of this research that may relate to my decision to continue participation, I will be informed. I have not given up any of my legal rights or released any individual or institution from liability for negligence.

The research team has adequately answered all the questions I have asked about this study, my participation, and the procedures involved. I understand that Ulf Ahlstrom or another member of the research team will be available to answer any questions concerning procedures throughout this study. If I have questions about this study or need to report any adverse effects from the research procedures, I will contact Ulf Ahlstrom at (609) 485-8642.

Discomfort and Risks:

In the part-task experiments, the screen may flicker back-and-forth between two images, at the rate of several times per second. For healthy individuals, there are no reported adverse effects of this common presentation technique. However, such flickering could cause seizures in epileptics. If you experience any discomfort due to the presentation of the images, please alert the experimenter immediately.

I agree to immediately report any injury or suspected adverse effect to Ulf Ahlstrom at (609) 485-8642.

Signature Lines:

I have read this informed consent form. I understand its contents, and I freely consent to participate in this study under the conditions described. I understand that, if I want to, I may have a copy of this form.

Research Participant:_____   Date:_____

Investigator:_____   Date:_____

Witness:_____   Date:_____

Appendix B
Research Staff List

Name	Role	Responsibility
Ahlstrom, Ulf	Test Lead	Lead developer of test plan, test conductor, data analysis.
Caddigan, Eamon	Human Factors Specialist	Co-developer of test plan, test conductor, data analysis.
Granich, Thomas	Software Engineer	Implements simulator system and cockpit data recordings.
Hallman, Kevin	Human Factors Specialist	Test conductor, data analysis
Johnson, Ian	WTIC Human Factors Lead	Coordinate on test plan and test effort, technical review of test products and deliverables.
Karaska, Kimberly	Software Engineer	Implements simulator system and cockpit data recordings.
Kukorlo, Matt	Pilot Subject Matter Expert	Flight Scenario Developer. Simulation SME.
Lyman, Stephen	Aerospace Engineer	Implements simulator flight models and cockpit out-the-window displays, data recordings.
Mauroff, Jim	Computer Scientist	Implements audio and video capabilities for the cockpits, data recordings.
Pokodner, Gary	WTIC Program Manager	Track project, conduct interim reviews, final acceptance of Deliverables.
Racine, Nicole	Human Factors Specialist	Co-developer of test plan, test conductor, data analysis.
Rehman, Al	Cockpit Simulator Laboratory Manager	Identify and provide subject pilots for study, update simulators and hardware to meet study requirements.
Schulz, Ken	Human Factors Specialist	Co-developer of test plan, test conductor, data analysis.
Truong, Cynthia	Software Engineer	Implements simulator system and cockpit data recordings.

Appendix C

Biographical Questionnaire

Instructions:
    This questionnaire is designed to obtain information about your background and experience as a pilot. Researchers will only use this information to describe the participants in this study as a group. Your identity will remain anonymous.

Demographic Information and Experience

1. What pilot certificate and ratings do you hold? (circle as many as apply)	Private Commercial ATP Glider SEL SEA MEL Airship Instrument CFI CFII MEI Helicopter A&P IA

2. What is your age?	_____ Years

3. Approximately how many actual total flights hours do you have?	_____ Hours

4. Approximately how many actual instrument hours do you have?	_____ Hours

5. Approximately how many instrument hours have you logged in the last 6 months (simulated and actual)?	_____ Hours

6. List all (if any) in-flight weather presentation systems you have used during a flight to make actual weather judgments? (not including onboard radar or Stormscope)?

    _____

    _____

    _____

7. Have you had any training in weather interpretation other than basic pilot training (for example, courses in meteorology)? If so, to what extent?

    _____

    _____

    _____

Thank you very much for participating in our study; we appreciate your help.

Mobile Device Proficiency Questionnaire (MDPQ-16)

About the MDPQ
This questionnaire asks about your ability to perform a number of tasks with a **mobile device**.

What is a Mobile Device?
A mobile device is a device that allows you to perform many of the same tasks as a standard computer but without the use of a physical keyboard and mouse. Instead, these devices use a touchscreen as their interface between the user and computer programs (called Apps – short for applications).

Mobile devices come in many sizes. Shown above are two different sized tablets and a smartphone. These are the types of devices we are interested in.

Please answer each question by placing an X in the box that is most appropriate.

If you have not tried to perform a task with a mobile device or do not know what a task is, please mark "NEVER TRIED", regardless of whether or not you think you may be able to perform the task. **Remember, you are rating your ability to perform each of these tasks specifically using a mobile device (tablet or smartphone).**

1. Mobile Device Basics

Using a mobile device I can:	Never tried (1)	Not at all (2)	Not very easily (3)	Somewhat easily (4)	Very easily (5)
a. Navigate onscreen menus using the touchscreen					
b. Use the onscreen keyboard to type					

2. Communication

Using a mobile device I can:	Never tried (1)	Not at all (2)	Not very easily (3)	Somewhat easily (4)	Very easily (5)
a. Send emails					
b. Send pictures by email					

3. Data and File Storage

Using a mobile device I can:	Never tried (1)	Not at all (2)	Not very easily (3)	Somewhat easily (4)	Very easily (5)
a. Transfer information (files such as music, pictures, documents) on my mobile device to my computer					
b. Transfer information (files such as music, pictures, documents) on my computer to my mobile device					

4. Internet

Using a mobile device I can:	Never tried (1)	Not at all (2)	Not very easily (3)	Somewhat easily (4)	Very easily (5)
a. Find information about my hobbies and interests on the Internet					
b. Find health information on the Internet					

5. Calendar

Using a mobile device I can:	Never tried (1)	Not at all (2)	Not very easily (3)	Somewhat easily (4)	Very easily (5)
a. Enter events and appointments into a calendar					
b. Check the date and time of upcoming and prior appointments					

6. Entertainment

Using a mobile device I can:	Never tried (1)	Not at all (2)	Not very easily (3)	Somewhat easily (4)	Very easily (5)
a. Use the device's online "store" to find games and other forms of entertainment (e.g. using Apple App Store or Google Play Store)					
b. Listen to music					

7. Privacy

Using a mobile device I can:	Never tried (1)	Not at all (2)	Not very easily (3)	Somewhat easily (4)	Very easily (5)
a. Setup a password to lock/unlock the device					
b. Erase all Internet browsing history and temporary files					

8. Troubleshooting & Software Management

Using a mobile device I can:	Never tried (1)	Not at all (2)	Not very easily (3)	Somewhat easily (4)	Very easily (5)
a. Update games and other applications					
b. Delete games and other applications					

Weather Questions

Please circle your answer

1. **What are characteristics of unstable air?**

A) Turbulence and good surface visibility.

B) Turbulence and poor surface visibility.

C) Nimbostratus clouds and good surface visibility.

2. **A temperature inversion would most likely result in which weather condition?**

A) Clouds with extensive vertical development above an inversion aloft.

B) Good visibility in the lower levels of the atmosphere and poor visibility above an inversion aloft.

C) An increase in temperature as altitude is increased.

3. **The amount of water vapor which air can hold depends on the**

A) dew point.

B) air temperature.

C) stability of the air.

4. **What clouds have the greatest turbulence?**

A) Towering cumulus.

B) Cumulonimbus.

C) Nimbostratus.

5. **In which meteorological environment is aircraft structural icing most likely to have the highest rate of accumulation?**

A) Cumulonimbus clouds.

B) High humidity and freezing temperature.

C) Freezing rain.

6. **For most effective use of the Radar Summary Chart during preflight planning, a pilot should**

A) consult the chart to determine more accurate measurements of freezing levels, cloud cover, and wind conditions between reporting stations.

B) compare it with the charts, reports, and forecasts to form a mental three-dimensional picture of clouds and precipitation.

C) utilize the chart as the only source of information regarding storms and hazardous conditions existing between reporting stations.

7. **What relationship exists between the winds at 2,000 feet above the surface and the surface winds?**

A) The winds at 2,000 feet and the surface winds flow in the same direction, but the surface winds are weaker due to friction.

B) The winds at 2,000 feet tend to parallel the isobars, while the surface winds cross the isobars at an angle toward lower pressure, and are weaker.

C) The surface winds tend to veer to the right of the winds at 2,000 feet, and are usually weaker.

8. **While flying a 3-degree glide slope, a headwind shears to a tailwind. Which conditions should the pilot expect on the glide slope?**

A) Airspeed and pitch attitude decrease, and there is a tendency to go below glide slope.

B) Airspeed and pitch attitude increase, and there is a tendency to go above glide slope.

C) airspeed and pitch attitude decrease, and there is a tendency to remain on the glide slope.

9. **The Hazardous Inflight Weather Advisory Service (HIWAS) is a continuous broadcast over selected VORS of**

A) SIGMETs, CONVECTIVE SIGMETs, AIRMETs, Severe Weather Forecast Alerts (AWW), and Center Weather Advisories (CWA).

B) SIGMETs, CONVECTIVE SIGMETs, AIRMETs, Wind Shear Advisories, and Severe Weather Forecast Alerts (AWW).

C) Wind Shear Advisories, Radar Weather Reports, SIGMETs, CONVECTIVE SIGMETs, AIRMETs, and Center Weather Advisories (CWA).

10. **If you fly into severe turbulence, which flight condition should you attempt to maintain?**

A) Constant airspeed (VA).

B) Level flight attitude.

C) Constant altitude and constant airspeed.

11. A pilot can expect a wind-shear zone in a temperature inversion whenever the windspeed at 2,000 to 4,000 feet above the surface is at least

A) 10 knots.

B) 15 knots.

C) 25 knots.

12. A high cloud is composed mostly of

A) ozone.

B) condensation nuclei.

C) ice crystals.

13. Where can wind shear associated with a thunderstorm be found? Choose the most complete answer.

A) In front of the thunderstorm cell (anvil side) and on the right side of the cell.

B) In front of the thunderstorm cell and directly under the cell.

C) On all sides of the thunderstorm cell and directly under the cell.

14. Maximum downdrafts in a microburst encounter may be as strong as

A) 8,000 feet per minute.

B) 7,000 feet per minute.

C) 6,000 feet per minute.

15. Which family of clouds is least likely to contribute to structural icing on an aircraft?

A) Low clouds.

B) High clouds.

C) Clouds with extensive vertical development.

16. The surface Analysis Chart depicts

A) actual pressure systems, frontal locations, cloud tops, and precipitation at the time shown on the chart.

B) frontal locations and expected movement, pressure centers, cloud coverage, and obstructions to vision at the time of chart transmission.

C) actual frontal positions, pressure patterns, temperature, dew point, wind, weather, and obstructions to vision at the valid time of the chart.

17. One weather phenomenon which will always occur when flying across a front is a change in the

A) wind direction.

B) type of precipitation.

C) stability of the air mass.

18. Which weather phenomenon signals the beginning of the mature stage of a thunderstorm?

A) The appearance of an anvil top.

B) Precipitation beginning to fall.

C) Maximum growth rate of the clouds.

19. What is an important characteristic of wind shear?

A) It is primarily associated with the lateral vortices generated by thunderstorms.

B) It usually exists only in the vicinity of thunderstorms, but may be found near a strong temperature inversion.

C) It may be associated with either a wind shift or a wind speed gradient at any level in the atmosphere.

20. Thrust is managed to maintain IAS, and glide slope is being flown. What characteristics should be observed when a headwind shears to be a constant tailwind?

A) PITCH ATTITUDE: Increases; REQUIRED THRUST: Increased, then reduced; VERTICAL SPEED: Increases; IAS: Increases, then decreases to approach speed.

B) PITCH ATTITUDE: Decreases; REQUIRED THRUST: Increased, then reduced; VERTICAL SPEED: Increases; IAS: Decreases, then increases to approach speed.

C) PITCH ATTITUDE: Increases; REQUIRED THRUST: Reduced, then increased; VERTICAL SPEED: Decreases; IAS: Decreases, then increases to approach speed.

Appendix F

Probe Questions

Probe Questions Administered at the 11, 20, and 35-Minute Marks of the Simulated Flight
Note: The questions designed to assess whether participants detected each METAR change are bolded.

t = 11 min (At *t* = 10 min the METAR at KMRB changed from VFR to IFR)
1. Please estimate your visibility at this point.
2. Have you checked in with ATC? If so, which ATC control facility did you first check in with?
3. At what altitude did you check in with ATC?
4. What ATC control facility were you handed off to?
5. After the East Texas VOR, what was your next navigational facility and what course did you set?
6. Were there any thunderstorms or other weather-related changes in the areas of Dulles and Martinsburg?

t = 20 min (At *t* = 19 min the METARs at KBWI, KDCA, KESN, KIAD, and KJST changed from VFR to IFR)
1. Please estimate your visibility at this point.
2. What ATC control facility are you communicating with?
3. What navigational facility are you using?
4. What is your current heading?
5. Did you notice any changes in the on-screen weather presentation since the last time the simulation was paused?

t = 35 min (At *t* = 30 min the METAR at KHGS changed from VFR to IFR)
1. Please estimate your visibility at this point.
2. What ATC control facility are you communicating with?
3. Did you notice any changes in the on-screen weather presentation since the last time the simulation was paused?
4. Did you notice the METAR status at HGR (Hagerstown)?
5. What is your plan of action?

Appendix G

1. Using the weather presentation, how easy was it to see the METAR information?

Very Hard				**Very Easy**
1	2	3	4	5

2. How easy was it to determine when a METAR symbol changed from VFR to IFR?

Very Hard				**Very Easy**
1	2	3	4	5

3. To what degree did the METAR information affect your decision to stay on your course or to fly to an alternate destination airport?

Very Hard				**Very Easy**
1	2	3	4	5

4. To what degree did the METAR information affect your decision to stay VFR or to request an IFR flight plan?

Very Hard				**Very Easy**
1	2	3	4	5

5. How would you rate the benefits of the weather presentation you used to other sources of weather information (ATIS, Flight Watch, etc.)?

Very Hard				**Very Easy**
1	2	3	4	5

6. How much do you think the weather presentation decreased your workload?

Very Hard				**Very Easy**
1	2	3	4	5

7. How much did you trust the weather presentation to give you correct information?

Very Hard				**Very Easy**
1	2	3	4	5

8. How easy was it to determine the position of the aircraft based on the presentation?

Very Hard				**Very Easy**
1	2	3	4	5

Thank you very much for participating in our study; we appreciate your help.

Appendix H

1. Using the weather presentation, how easy was it to see the distance from the aircraft to precipitation cells?

Very Hard								**Very Easy**
1	2	3	4	5	6	7	8	9

2. At the beginning of the scenario, did you fly closer than 20 nautical miles (nmi) from the precipitation cells?

……☐ Yes……………☐ No ……

3. How close to the precipitation cells did you get?

……☐ 20-25 nmi……☐ 15-20 nmi……☐ 10-15 nmi ……☐ less than 10 nmi

4. How would you rate the benefits of the *blue line* reminder of the distance from the aircraft to precipitation cells?

No Benefit								**Very Beneficial**
1	2	3	4	5	6	7	8	9

5. During the scenario flight, the *blue line* reminder distance was set to 20 nmi. If you would use this implementation during actual flights – what nmi reminder distance would you set?

☐ 30 nmi or more…☐ 20-25 nmi…☐ 15-20 nmi…☐ 10-15 nmi …☐ less than 10nmi

6. The current FAA recommendation is to stay 20 nmi away from storms. Do you agree with this recommendation?

……☐ Yes……………☐ No ……

7. If you would determine the recommendation for a distance to storms, what distance (in nmi) would you pick?

☐ 30 or more…..☐ 20 (the current distance) …..☐ 15 ……...☐ 10 ……...☐ less than 10

8. To what degree did the weather display information affect your decision to stay on your course or to fly to an alternate destination airport?

Very Little								**Very Much**
1	2	3	4	5	6	7	8	9

9. How would you rate the benefits of the weather presentation you used to other sources of weather information (ASOS, Flight Watch, etc.)?

Very Low								**Very High**
1	2	3	4	5	6	7	8	9

10. How much do you think the weather presentation decreased your cognitive workload (i.e., it gave you easy access to information that you otherwise would have to get from other sources)?

Very Little								**Very Much**
1	2	3	4	5	6	7	8	9

11. How much did you trust the weather presentation to give you correct information?

Very Little								**Very Much**
1	2	3	4	5	6	7	8	9

12. Did the precipitation cells appear closer in the out-the-window (OTW) view, or on the weather display?

OTW				**Neither**				**Display**
1	2	3	4	5	6	7	8	9

13. Did the precipitation cells appear more intense in the out-the-window view, or on the weather display?

OTW				**Neither**				**Display**
1	2	3	4	5	6	7	8	9

14. Did the weather situation appear to change more rapidly in the out-the-window view, or on the weather display?

OTW				**Neither**				**Display**
1	2	3	4	5	6	7	8	9

15. During flight, how easy was it to determine the visibility (in nmi) from the out-the-window view?

Very Hard								**Very Easy**
1	2	3	4	5	6	7	8	9

16. During flight, did you ever notice that the visibility was decreasing?

……☐ Yes……………☐ No ……

17. In your estimate, what was the lowest visibility that you encountered during flight?

…☐ 5-6 nmi …☐ 4-5 nmi …☐ 3-4 nmi…☐ 2-3 nmi…☐ 1-2 nmi …☐ less than 1 nmi

18. How would you rate the benefits of the *blue line* reminder to areas of forecasted visibility of 1-3 nmi?

No Benefit								**Very Beneficial**
1	2	3	4	5	6	7	8	9

19. During the scenario flight, the *blue line* reminder distance to forecasted areas of 1-3 nmi visibility was set to 20 nmi. If you would use this implementation during actual flights – what nmi reminder distance would you set?

    ☐ 30 nmi or more…☐ 20-25 nmi…☐ 15-20 nmi…☐ 10-15 nmi …☐ less than 10nmi

20. To what degree did the *blue line* reminder of visibility conditions affect your decision to stay on your course or to fly to an alternate destination airport?

Very Little								**Very Much**
1	2	3	4	5	6	7	8	9

21. During the scenario, did you ever use ASOS information?

    ……☐ Yes……………☐ No ……

Thank you very much for participating in our study; we appreciate your help.